OUTLOOK
Colorectal cancer

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

SCIENCE IN INDIA

A special issue explores India's grand challenges PAGE 141

AUTOBIOGRAPHY

A LIFE IN THE FAST LANE

Oliver Sacks's journey from biker to thinker

PAGE 158

NEUROSCIENCE

INSECTS KNOW THEIR PLACE

The brains behind 'gnat nav'

PAGES 165 & 186

ASTROPHYSICS

TURN OFF THE GAS

How star formation is blocked in local galaxies

PAGES 164 & 192

NATURE.COM/NATURE
14 May 2015 £10
Vol. 521, No. 7551
9 770028 083095 20

THIS WEEK

A nation with ambition

India is making great strides in improving its science, but it needs to look carefully at its approach and work with the rest of the world if it is to realize its full potential.

The website of the Indian Department of Science and Technology proudly states that "India is one of the top-ranking countries in the field of basic research".

It is true that India has made considerable progress in areas such as biotechnology, renewable energy and aerospace. But it is also mired in deep problems that impede innovation and are hampering the country's progress. India has a puny scientific workforce, relatively few high-quality universities, an anaemic manufacturing sector and an epidemic of red tape. The result is that many Indian scientists head to other countries for training and jobs.

It would be easy to argue that a lack of funding is holding India back and stopping it from becoming a science superpower. The country devotes less than 1% of its gross domestic product to research and development, which puts it far behind emerging nations such as China and Brazil, as well as the established economies of the United States and Europe. But more money will not cure India's multiple science ills, as *Nature* documents this week in a special issue on the state of research in the country (see page 141).

One of India's biggest challenges is to boost its science to help drive national development. As a start, it must expand its research workforce. But that will require more high-quality universities and appropriate jobs for their graduates. The government is taking steps in the right direction. It has established tax incentives for research and development that are among the best in the world. These have helped to boost research investments by a few industries, but have yet to drive widespread innovation.

In tandem, India must tackle the bureaucratic morass that is impeding research and innovation. Scientists complain that funds for grants routinely arrive months late and that it can take years to fill positions. As a measure of the problem, one-third of the national laboratories, which are overseen by the prestigious Council of Scientific and Industrial Research (CSIR), lack permanent leaders (see page 144). Even the CSIR is run by a temporary director-general, Madhukar Garg, who told *Nature* that if the organization continues along these lines, "it will affect the national innovation system as a whole".

Prime Minister Narendra Modi, like his predecessors, has denounced the bureaucratic brakes holding back science, but there has been little progress here. A key to solving the issue is to elevate talented scientists who have administration experience into positions of responsibility. One example is Krishnaswamy VijayRaghavan, who is profiled on page 148. He is a gifted geneticist who in 2013 took over as head of the Department of Biotechnology, the leading funder of bioscience research grants. Among other changes, he is attempting is to streamline the notoriously cumbersome grant-application process.

India could use some help. Compared with some other developing nations, it has a relatively low level of international collaboration, even with the United Kingdom, with which it shares a unique history. It bodes well that the new UK minister of universities and science, Jo Johnson, has a strong interest in India. In fact, he co-edited a book entitled *Reconnecting Britain and India: Ideas for an Enhanced Partnership* (Academic Foundation, 2012).

India does, however, need to look closely at the changes it is making, because not all are positive. As part of its effort to encourage development, the Modi administration has tried to silence some critics of policies on energy, climate and human rights. In April, the Indian government revoked the registrations of thousands of non-governmental organizations (NGOs) that receive foreign funds, and it has frozen the assets of Greenpeace over claims that it had violated reporting rules about foreign contributions. On 6 May, the US ambassador to India, Richard Verma, warned about "the potentially chilling effects of these regulatory steps focused on NGOs".

"The solution is not to silence discussion or to shrink environmental oversight."

Some scientists might be tempted to applaud India's clampdown on environmental groups, which have stymied certain research initiatives. In March, environmentalists held up construction of a major neutrino observatory with debatable claims that the facility would harm an aquifer. And *Nature* reports this week that the Modi government has quietly moved forward with trials of genetically modified crops, which have long been desired by biotech researchers but have been impeded by environmental groups (see page 138).

But scientists in India should not cheer the government's attempts to suppress dissent, even if it helps them to achieve their research goals. It would be wrong to blame environmental advocates for India's lengthy and fault-ridden procedures for weighing up the impact of projects. The solution is not to silence discussion or to shrink environmental oversight. Rather, India should make strategic improvements to the environmental evaluation process that balance progress with protection. ■

Challenging times

A European initiative to ban animal research has galvanized resistance.

The Stop Vivisection initiative has been panicking European researchers since it was first proposed in 2012, but its long-trailed public hearing this week at the European Parliament in Brussels turned out to be a pretty grey affair.

The duo who presented the initiative — which calls for the replacement of the 2010 European Union (EU) directive on the use of animals in scientific research with legislation banning all animal research — spoke calmly but unconvincingly. Their extremist claims, that animal models have no predictive value for human disease, drew thin and only

occasional ripples of agreement from a cluster of supporters seated at the back of the half-filled auditorium.

The hearing was part of the EU's move to expand direct democracy by introducing European Citizens' Initiatives, which allow individuals to launch requests for legislation. A proposed initiative that collects at least one million signatures from at least seven EU countries wins the right to a public hearing in the Brussels parliament and obliges the European Commission to consider whether new legislation is warranted.

As Kay Davies, an animal researcher at the University of Oxford, UK, wrote in *Nature* last week, in this case, it is not (see *Nature* **521,** 7; 2015).

For too long, activists have been left to dominate animal-research debates in many European countries. Their frequently inaccurate declarations — along with their not-infrequent physical attacks and death threats — have gone largely unchallenged by the scientific community and by the agencies and politicians who support the community's work. This has been slowly changing in the past few years, mostly thanks to the efforts of UK-based scientists and science organizations, who have emerged from their bunkers to set the record straight.

Germany, despite its status as one of Europe's major scientific powerhouses, has lagged behind in this effort. But a recent incident has sparked a remarkable change — one that should shore up support to protect the EU directive.

One of the country's top neuroscientists, Nikos Logothetis, a director at the Max Planck Institute for Biological Cybernetics in Tübingen, last month gave up a long and painful struggle to maintain his primate laboratory, which had been targeted by animal activists. Unable to handle the death threats and insults to himself and his family, on 22 April he told local authorities who handle licences for animal experimentation that he would wind down his primate work and continue his work on rats only. This would mean reducing the scope of his research questions to levels still valuable for understanding general principles of neuronal action, but no longer directly translatable to human investigations.

Logothetis's problems began last September, when a German television channel aired a documentary using footage of his macaque monkeys secretly filmed by an animal-activist infiltrator. It seemed to show maltreatment of the animals. The resulting scandal led to a series of investigations that exonerated him and suggested that the behaviour of the monkeys had been staged for the camera. A police investigation is still going on.

His decision to quit primate work dismayed many of his colleagues, and so did its timing. Coming so close to this week's public hearing, they feared that it would be presented as a victory for the Stop Vivisection proponents. But something quite different happened: a swell of support for Logothetis and the type of primate research he carries out.

First, politicians at the highest levels reacted with unprecedented speed and clarity to mount an unambiguous defence of the scientific use of non-human primate research. The research minister in the state government of Baden-Württemberg, where Tübingen is located, condemned as cynical and exploitative the wild claims that Logothetis's decision implied that research with monkeys was not after all necessary. The federal research minister stated that such research was still crucial for developing treatments for brain disorders such as dementias.

"For too long, activists have been left to dominate animal-research debates in many European countries."

And a new policy of the Max Planck Society to be more open about its animal research showed its teeth. On 30 April, the society released a statement of regret about Logothetis's decision and confirmed its own commitment to continue supporting research using non-human primates. The society's president, Martin Stratmann, a materials scientist who took office last year and who has selected the handling of the animal issue as a priority for 2015, spoke out to pledge greater protection of its researchers against attack. At the grass roots, colleagues in Tübingen launched a petition to support Logothetis that has received more than 4,000 signatures from scientists around the world.

This outspoken support has been echoed elsewhere. Parliamentary debates on animal research in Italy this week, where animal groups have been particularly active in the past few years, questioned rather than accepted animal-activist claims. Sixteen European Nobel laureates published an open letter in UK and German newspapers to rebut the Stop Vivisection campaign, joining a similar statement by 149 major research organizations and patient groups.

The European Parliament has until 3 June to decide what to do. It should listen to the loud and unified voice of the continent's scientists, and then do precisely nothing. ■

Polls apart

The UK voter opinion polls show that an anomalous answer can be the correct one.

Britain's new Conservative government has barely settled into office, but already the results of last week's general election have got certain members of UK society fearing for their future. They are scorned by the tabloid press and social media; even serious observers are questioning whether the country has been in thrall to them for too long. An inquiry has already been announced.

Opinion pollsters, the media told everyone, were predicting the closest election for decades. Labour and the Conservatives were neck and neck; weeks of constitutional chaos would follow the election as mandarins and officials wrestled with competing and overlapping political claims to power. The small print says that opinion polls should always be taken with a decent pinch of salt. But who reads the small print when there is an election on and a 24-hour news cycle to fill?

It took a single poll of voters post-voting to reveal the truth, which was confirmed as the counted results flooded in: David Cameron's Conservative Party had grabbed 37% of the vote (see page 134). That was nearly seven percentage points ahead of Labour and, crucially, well outside the margins of error of all the previous deadlocked polls.

Amid the fallout, a single polling firm revealed that it had correctly predicted — and then buried — the result. Gathered the day before the election, its poll results seemed so out of line with what everyone else was saying that the firm did not dare to publish them. "I chickened out of publishing the figures," confessed Damian Lyons Lowe, the chief executive of Survation in London. "Something I'm sure I'll always regret."

Nature's readers can surely sympathize. The question of how to deal with anomalous data is a centrepiece of research, and the results can make or break careers — or launch scientific revolutions. From the discovery of the ozone hole over Antarctica to the observation that some people seemed unaffected by HIV infection, unusual results — data that make you go 'hmmm' — have led scientists to question their methods, their knowledge and, ultimately, their understanding of the world.

The importance of anomalies in science has spawned its own sub-field of research into how researchers respond to them. In the mid-1980s, psychologists supported by US military funds went as far as constructing a bespoke computer program to recreate how Hans Krebs reacted to surprising results during his discovery of the urea cycle in 1932. Others conduct *in vivo* studies by filming astronomers and physicists as they wrestle with unexpected findings.

The ultimate test of anomalous data is, of course, to repeat the experiment. But that demands that scientists have the courage and insight to treat such results seriously in the first place. How many potential discoveries lie in the waste-paper bin of history because the cautious chickened out? ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

MADELEINE P. BALL

Regulate gene editing in wild animals

The use of genome-modification tools in wild species must be properly governed to avoid irreversible damage to ecosystems, says **Jeantine Lunshof**.

Gene editing is a hot topic following a flurry of interest in the use of CRISPR tools to modify human embryos. As an ethicist in a genome-engineering lab, I am an eyewitness to these recent scientific developments and I do have concerns about the way gene editing could be used. But they are not the concerns you might expect.

The ethical issues raised by human germline engineering are not new. They deserve consideration, but outcry over designer babies and precision gene therapy should not blind us to a much more pressing problem: the increasing use of CRISPR to edit the genomes of wild animal populations. Unless properly regulated and contained, this research has the potential to rapidly alter ecosystems in irreversible and damaging ways.

Scientists have already used CRISPR to modify mosquitoes and the fruit fly *Drosophila melanogaster*. And in combination with another molecular-biology technique called gene drive, they have found a way to massively increase the efficiency of spreading these transformations to offspring and through the population. Once introduced, these genetic changes are self-propagating. If released beyond the laboratory, the effects would spread with every new generation and would quickly run out of control.

Gene drive achieves rapid changes in a sexually reproducing population because it relies on genes that are capable of preferential spread through generations. Without this, introduced traits meet the statistical obstacle of Mendelian inheritance and take hold in a population much more slowly. Altering wild animal populations using gene drive aims to rapidly disrupt a particular trait, such as the ability of *Anopheles* mosquitoes to transmit malaria. It makes only a small-scale initial change to the relevant ecosystem and, in this example, the preliminary disruption would be restricted to the mosquito's natural habitat. But the risk of broader ecosystem disruption is unknown and would require extensive mathematical modelling to estimate.

The gene-drive technique was developed in our lab, and in the initial publication of the method, my colleagues called for strict and validated biosafety measures and public review and consent (K. M. Esvelt *et al. eLife* **3,** e03401; 2015). Meanwhile, work that combines CRISPR and gene-drive techniques is marching on. In what they call a "mutagenic chain reaction", scientists at the University of California, San Diego, have used the combined approach to alter *D. melanogaster* (V. M. Gantz and E. Bier *Science* **348,** 442–444; 2015). To me and others, this research raises serious and significant fears about biosafety. The work was done in a lab, but should any of the modified insects escape, they would be able to spread widely — unlike mosquitoes, which rely on ecological niches — and breed with the wild population. Experiments such as these should certainly be allowed, but only under the strictest conditions and with appropriate safeguards.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/igwpmu

In less than three years, CRISPR has become a key tool for biologists. 'Should they stop before it is too late?' is therefore an immaterial question. Careful assessment of the various applications of CRISPR shows that there is no single or universal ethical evaluation that could cover all of them. The different consequences of human genome modification in either somatic cells or the germ line, and the modification of the ecosystem through gene drives, call for different ethical and policy evaluations.

Some critics argue that the unpredictable effects that human germline genome editing could have on future generations make it dangerous and ethically unacceptable. Uncertainty, however, is not a useful way to judge ethical acceptability. Others highlight the potential non-therapeutic purposes of germline modification. From the standpoint of ethics, it is not clear why trait modification is by definition a bad thing. Moreover, the criteria for what is therapy and what is 'enhancement' are fluid.

WE NEED AN URGENT REVIEW OF BIOSAFETY PROTOCOLS FOR CRISPR AND GENE-DRIVE EXPERIMENTS IN WILD ORGANISMS.

The consequences of modifying human genomes will be limited because effects will always be restricted to humans — the index person and their line of descendants. Biosafety and biosecurity risks are not apparent at this moment. Regulation, if and when it comes, might need to be adapted to the local situation, to existing legislation and to cultural and religious normative frameworks.

Presented in those terms, the human applications of CRISPR are much less troubling than the possibility of ecosystem modification. By definition, such disruption has more severe, complex, system-level consequences, and the breadth of its impact and the duration of its effects are hard to model. Gene drives are designed to be reversible, but this still needs to be tested. Self-propagating modified organisms cannot be contained within national borders and pose major challenges for regulation and governance. We therefore need an urgent review of biosafety and biosecurity protocols for experiments — both in the lab and in field-scale trials — that combine CRISPR and gene-drive techniques in wild organisms. Funders and institutions must lay out and enforce regulations.

The work with gene editing has thrown a useful spotlight on these bioengineering tools. But from an ethical perspective, the question we should ask is not what CRISPR can do for humans, but what humans can do with CRISPR. ■

Jeantine Lunshof *is assistant professor at the University of Groningen, the Netherlands, and a visiting scientist at Harvard Medical School, Boston, Massachusetts, USA.*
e-mail: jelunshof@genetics.med.harvard.edu

RESEARCH HIGHLIGHTS

Selections from the scientific literature

CHEMISTRY

Crystal harbours molecular shuttle

A ring-shaped molecule that hops between two sites inside a porous crystal is the first molecular shuttle to operate in a solid-state material.

Molecular shuttles could one day act as switches to store data if they are held in a well-ordered array. A team led by Stephen Loeb and Robert Schurko at the University of Windsor, Canada, built their shuttle inside a metal–organic framework (MOF): a crystalline scaffold made from metal-containing nodes connected by carbon-based struts. These struts bore a circular rotaxane molecule that moved back and forth 283 times per second at room temperature.

The estimated high density of shuttles in the MOF shows the potential for enormous data capacity if such switches can be controlled, the authors say.

Nature Chem. **http://doi.org/4f2 (2015)**

MICROBIOLOGY

Gut biofilms could spur cancer

Chemicals secreted by gut bacteria are linked to human colon cancers.

Metabolites called polyamines are made by gut bacteria to help them to form sticky aggregates called biofilms, and are used by human cells to regulate proliferation. Cynthia Sears at Johns Hopkins University in Baltimore, Maryland, Gary Siuzdak at the Scripps Research Institute in La Jolla, California, and their colleagues compared tissue samples from human colon cancers to those from healthy people, both with and without biofilms.

They found that cancer tissue with biofilms had 62 times more of the polyamine metabolite N^1,N^{12}-diacetylspermine than did healthy tissue with biofilms. Yet in samples that were biofilm-free, the cancer tissue contained only around 7 times more polyamine than the healthy sample. Antibiotic treatment reduced levels of this metabolite, suggesting that it comes from bacteria.

Therapies that target polyamine formation and biofilms could be a way to treat colon cancer, the authors note.

Cell Metab. **http://doi.org/4jz (2015)**

WANG *ET AL., NATURE COMMUN.*

PALAEONTOLOGY

Early modern bird

Two newly discovered fossils from China are the oldest relative of modern birds found so far, and had relatively few traits of their dinosaur ancestors.

Min Wang and Zhonghe Zhou at the Institute of Vertebrate Paleontology and Paleoanthropology in Beijing and their colleagues describe two fossils from northeast China dating to 130.7 million years ago. The new species, *Archaeornithura meemannae* (pictured), boasted elaborate plumage — including a fan-shaped tail — and other defining features of modern birds such as a wishbone. This suggests that even these early birds, which are six million years older than previously discovered bird fossils, had already shed key dinosaur traits.

A. meemannae belongs to a family of wading birds, hinting that modern birds originated near water, the team says.

Nature Commun. 6, **6987 (2015)**

CLIMATE SCIENCE

Growing extremes in California rains

California's ongoing drought is a result of natural variability, at least for the 2013–14 period. But the state could see larger swings in wet and dry seasons by the end of this century owing to climate change.

Neil Berg and Alex Hall of the University of California, Los Angeles, used 34 climate models to study how precipitation extremes might change in California. They expect that between 2060 and 2100, the normally wet winter will be extremely dry twice as often as today, and extremely wet three times as often. The fluctuations could raise the risk of drought and flooding.

These changes could push the state's water supply to its limit.

J. Clim. **http://doi.org/4fz (2015)**

EVOLUTION

Bird beak to dinosaur snout

Chicken embryos with dinosaur-like faces provide clues as to how bird beaks evolved from dinosaur snouts.

Early in bird evolution, the twin bones that form the snout in dinosaurs and reptiles — the premaxillae — grew longer and joined together, eventually forming the beak. Bhart-Anjan Bhullar, now at the University of Chicago in Illinois, Arhat Abzhanov at Harvard University in Cambridge, Massachusetts, and their colleagues analysed the development of beaks in embryonic chickens and emus, and snout development in reptiles such as alligators.

They found that two proteins involved in facial development, FGF and Wnt, might have a role in beak evolution. When

they inhibited these genes in developing chicken eggs — making the expression pattern more like that of reptiles and other vertebrates — the premaxillae were separate and much shorter in some chicken embryos than in others. (None was allowed to hatch.)

X-ray scans of the embryonic skulls showed that they more closely resembled the bones of early birds and dinosaurs than those of modern chickens.
Evolution **http://dx.doi.org/10.1111/evo.12684 (2015)**

ENTOMOLOGY

Ants dig differently depending on dirt

Fire ants adjust their digging behaviour when building underground nests, depending on the size and dampness of the material they are working with.

Daniel Goldman and his colleagues at the Georgia Institute of Technology in Atlanta studied fire ants (*Solenopsis invicta*; **pictured**) as they constructed tunnel networks in silica particles designed to mimic soil of varying moisture levels and particle size. Using X-ray imaging, the team found that up to a point, ants dug deeper tunnels through clay-like and fine-grained particles as the moisture content increased, using the growing stickiness of the particles. When digging, the ants either carried individual grains away if they were large, or gathered and compressed smaller particles into a pellet.

The authors say that this adaptability could help to explain how this species has successfully colonized large parts of North America.
J. Exp. Biol. 218, **1295–1305 (2015)**

CAMILA DOS SANTOS, COLD SPRING HARBOR LAB

DANIEL GOLDMAN, DARIA MONAENKOVA

ASTROPHYSICS

Farthest galaxy measured

Astronomers have observed a distant galaxy as it looked just 650 million years after the Big Bang, making it the farthest galaxy to have its distance reliably measured.

Pascal Oesch at Yale University in New Haven, Connecticut, and his colleagues used a telescope at the W. M. Keck Observatory in Hawaii to study the galaxy EGS-zs8-1. To gauge its age, they measured how much its light had been stretched, or redshifted, as the light travelled across the expanding Universe. They found that the object is surprisingly bright and massive for such a young galaxy — a sign of rapid star formation.

Their analysis adds weight to a theory that the peculiar colours of early galaxies, compared to later objects, may be the result of interaction between these rapidly forming stars and the gas around them.
Astrophys. J. Lett. 804, **L30 (2015)**

INFECTIOUS DISEASE

Silenced gene keeps malaria out

Researchers have uncovered a possible target for anti-malarial therapies by suppressing gene expression in blood stem cells.

Malaria-causing parasites attack mature red blood cells, which lack DNA, making it hard to test which genes make the cells vulnerable to malaria infection. Manoj Duraisingh at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, and his colleagues have sidestepped this problem by studying the stem cells that develop into red blood cells and contain DNA.

They used RNA molecules to knock down gene activity in the stem cells, induced them to develop into red blood cells and then exposed them to *Plasmodium falciparum* — the most deadly malaria pathogen.

They found that the parasite binds to a cell-surface protein called CD55 and could not infect cells that did not produce it. Red blood cells can function normally without CD55, suggesting that the protein could be a therapeutic target.
Science 348, **711–714 (2015)**

SOCIAL SELECTION

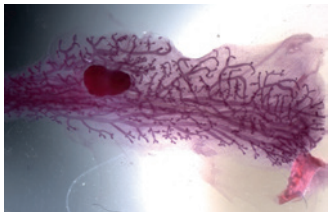
Popular articles on social media

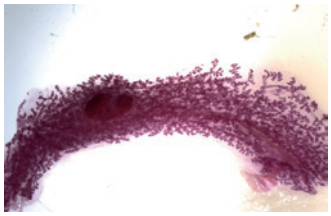
Mystery of telescope signals solved

A report on the surprising origins of rogue signals picked up by a radio telescope has been simmering on social media. After more than four years, researchers using the Parkes radio telescope in New South Wales, Australia, have identified the source of the mysterious signals: a microwave oven in the facility's break room. The news quickly spread on Twitter. "And the result of the story: don't use microwaves next to radio telescopes!" tweeted Karina Voggel, an astronomy PhD student at the European Southern Observatory in Garching, Germany. "A nice reminder that not all radio transients are microwave ovens — even if some are," tweeted Chris Lintott, an astrophysicist at the University of Oxford, UK.
http://arXiv.org/abs/1504.02165v1 (2015)

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/o7mh1p





EPIGENETICS

Mammary cells have a memory

Specific DNA changes make the mammary glands in pregnant mice gear up for milk production faster if the animals have been pregnant before.

Many women report that breast-feeding becomes easier with each pregnancy. To find out why, Gregory Hannon at Cold Spring Harbor Laboratory in New York and his colleagues gave pregnancy hormones to female mice and monitored changes in their mammary glands over time.

In mice that had been pregnant once before, the milk-producing ducts expanded in number more quickly (**pictured**, right) and generated milk proteins sooner than the ducts of mice that were first-time mothers (pictured, left). Genetic analysis of mammary-gland cells after the animals had stopped lactating showed that a previous pregnancy resulted in a long-term loss of methyl groups from DNA in genes that are activated during lactation.

This epigenetic 'memory' primes these genes for fast reactivation in a subsequent pregnancy, the authors say.
Cell Rep. **http://doi.org/4j2 (2015)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

EVENTS

UK election impact

The incumbent Prime Minister David Cameron and his Conservative Party won the United Kingdom's election on 7 May with a surprise outright majority that flummoxed pollsters. The win has implications for research because ministers are likely to emphasize austerity allied to economic growth — a pillar of the previous Conservative-led coalition government. There are no promises to protect science funding, which has been frozen for five years. The Conservative pledge to hold a 2017 referendum on Britain withdrawing from the European Union could adversely affect UK science if the country leaves. Jo Johnson, a modern-history graduate, former banker and journalist, has been appointed as Minister for Universities and Science. See page 134 for more.

JAMES GIAHYUE/REUTERS

Liberia celebrates becoming Ebola free

Liberia has become the first of the three main countries affected by Ebola to be declared officially free of the disease, ending its 15-month epidemic. The last person in Liberia known to have Ebola died on 27 March and was buried the next day. This means that, as of 9 May, the country has had no new cases for 42 days (twice the maximum incubation time) since the last burial — the criterion used by the World Health Organization (WHO) for declaring a country Ebola free. But with cases continuing in Sierra Leone and Guinea, the WHO has warned against complacency. See go.nature.com/mhimlt for more.

RAY BARKER

Nature's viability and health is as much dependent on its business leaders as on its editors. It is with sorrow that we record the death of Ray Barker (1936–2015), who was managing director of Macmillan Magazines from 1991 to 1999. He had a deep knowledge of publishing, and as publisher of *Nature* he increased financial robustness following a period of weakness. He also oversaw substantial growth of the stable of Nature research journals. His colleagues remember his exceptional ability to recognize their contributions; according to Stefan von Holtzbrinck, of the publishing family that bought Macmillan during Ray's tenure, "Many of the company's achievements stemmed from Ray's unique gift to spot and train extraordinary talents. His charisma inspired everyone to do the best possible job and go the last mile."

Arab Mars mission

The United Arab Emirates plans to launch a spacecraft to Mars in 2020, it announced on 6 May. The probe, named Hope, will be the first Arab spacecraft to go to the red planet. It will aim to model Mars's atmosphere from an elliptical orbit, studying temperature, gases and weather phenomena on each pass. It will be managed by the UAE Space Agency and controlled from the Mohammed bin Rashid Space Centre in Dubai. Partners in the mission include the University of Colorado Boulder, the University of California, Berkeley, and Arizona State University in Tempe.

BUSINESS

Mega-merger bid

Monsanto announced on 8 May that it had attempted to buy another leading agricultural-technology company, Syngenta of Basel, Switzerland. The offer — 449 Swiss francs (US$481) per Syngenta share, or about US$45 billion in total — was rejected, with Syngenta arguing that Monsanto, based in St Louis, Missouri, had undervalued the firm. If the purchase were to be successful eventually, it would unite Syngenta's business, a leader in pesticides and herbicides, with Monsanto's strong seed business, which trades in both conventionally bred and genetically engineered crops.

Drug-use panel

Johnson & Johnson has set up an independent bioethics panel to decide whether to grant patients access to experimental medicines, the New Jersey drug firm said on 7 May. The panel will be chaired by Arthur Caplan, a bioethicist at New York University. It will operate on a trial basis to consider compassionate-use requests for a single, experimental drug that is being developed by the firm's subsidiary Janssen. More than 20 US states are considering 'right to try' laws that would encourage drug companies to

offer investigational medicines to patients. See go.nature.com/84wob6 for more.

Rare-disease deal

Alexion Pharmaceuticals in Cheshire, Connecticut, has agreed to pay US$8.4 billion to buy Synageva BioPharma in Lexington, Massachusetts. Synageva is developing a drug for a rare genetic disease called lysosomal acid lipase deficiency; the therapy is in late-stage clinical trials. The deal, announced on 7 May, highlights the pharmaceutical industry's growing interest in treatments for rare conditions, and will expand Alexion's repertoire of therapies for metabolic diseases.

TECHNOLOGY

Escape-pod test

Private spaceflight company SpaceX passed a crucial test by successfully deploying the emergency escape pod of its Crew Dragon rocket system. The capsule launched (**pictured**) from Cape Canaveral, Florida, on 6 May. As planned, it flew 1.5 kilometres above the Atlantic Ocean before the escape pod detached, unfurling three parachutes to splash down safely into the sea. Sensors aboard the pod showed that the conditions would have been safe for humans. SpaceX,

SPACEX

which is headquartered in Hawthorne, California, is expected to start shuttling astronauts into space in 2017.

POLICY

Disease monikers

The World Health Organization (WHO) has issued guidelines for naming new diseases. The rules, released on 8 May, recommend naming diseases according to what they do rather than after the individual, region or species in which they were found. The name swine flu, for instance, has hurt the pork industry even though the virus is not specific to pigs, and Middle Eastern countries have been concerned that the name of the viral infection Middle East Respiratory Syndrome will lead to stigmatization. Researchers should also avoid names with frightening words such as 'fatal', the WHO said.

RESEARCH

Symbolic CO_2 peak

The global monthly concentration of atmospheric carbon dioxide has exceeded 400 parts per million (p.p.m.) for the first time since agencies began to track it in the 1960s. Analysing air samples from 40 sites worldwide, the US National Oceanic and Atmospheric Administration calculated an average CO_2 level of 400.83 p.p.m. for March. Calibrated data for April are not yet available, but scientists expect CO_2 concentration to remain above 400 p.p.m. for April and throughout May, when it normally peaks. Daily levels first surpassed 400 p.p.m. in 2012 at Arctic sites.

Tailored tests

Personalized screening and medicines could save the United States billions of dollars, but such interventions are unlikely to emerge because of the structure of its health-care system, says a panel of doctors and economists in *The Lancet* (V. J. Dzau *et al. Lancet* http://doi.org/4km; 2015). Modelling by the group, led by Victor Dzau of the Institute of Medicine in Washington DC, found a 10% fall in diabetes and cancer cases as a result of using tailored tests would generate US$166 billion in improved health over 50 years. But the group notes that private insurers cover therapies by judging only short-term gains.

High-energy ties

The United States and CERN, Europe's high-energy physics lab near Geneva, Switzerland, pledged to align their long-term strategies for particle physics on 6 May. Although the United States has only observer status in CERN, it is a major contributor to experiments such as those at the Large Hadron Collider. CERN will now reciprocate with more-direct involvement in US-based experiments, notably in neutrino science. Also on 6 May, Turkey formalized its associate membership of CERN, which grants it access to CERN Council meetings.

COMING UP

16–20 MAY
Scientists and health-care professionals meet in Dublin for the European Congress of Endocrinology to discuss the latest advances in the field.
go.nature.com/p5mtoq

18–22 MAY
The latest research in the science of sound is presented at the Acoustical Society of America's spring meeting in Pittsburgh, Pennsylvania.
go.nature.com/wseczh

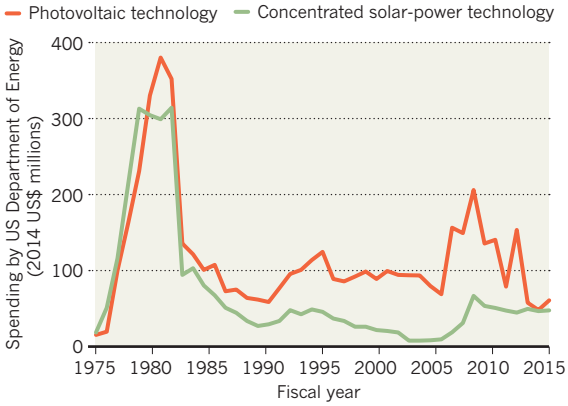
18–26 MAY
The World Health Organization holds its general assembly in Geneva, Switzerland. Delegates from all its member states discuss and decide on health policies, programmes and budgets.
go.nature.com/klqaym

TREND WATCH

The US Department of Energy should create a sustained and predictable programme of funding focused on new solar technologies, to support a "massive expansion" of solar power by mid-century, says a report from the Massachusetts Institute of Technology in Cambridge (see go.nature.com/ogv7wa). Released on 5 May, it urges policies to promote solar energy, including subsidies for generating technologies and incentivized programmes for developing low-cost technologies.

SOLAR SUPPORT

The US government has a long history of funding solar-energy technologies, but will need to raise investment in the future.



SOURCE: *THE FUTURE OF SOLAR ENERGY* (MIT, 2015)

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

PRABIN POKHREL/NEPAL POLICE/HANDOUT/EPA

The village of Langtang was buried by a landslide last month after a major earthquake struck Nepal.

EARTH SCIENCE

Mappers rush to pinpoint landslide risk in Nepal

Geologists say hazard posed by earthquake-loosened earth could linger for years.

BY ALEXANDRA WITZE

As Nepal digs out from the devastating magnitude-7.8 earthquake of 25 April, researchers are braced for the next geological hazard. In June, monsoon rains will begin to drench the hillsides destabilized by the quake, raising the risk of disastrous landslides.

Geologists are racing to identify areas that could collapse and bury villages or block important roads. They say that the hazard is likely to linger for years.

So far, the situation has been grim but better than expected. Scientists have identified thousands of landslides in satellite imagery, rather than the tens or hundreds of thousands they anticipated from the aftermaths of other large earthquakes in mountainous areas. "Something slightly mysterious has happened," says David Petley, a geoscientist at the University of East Anglia in Norwich, UK. "The landslide problem is far from trivial, but it's not as desperately serious as we might have feared."

The most violent event occurred in the Langtang valley, a popular trekking area in the mountains north of Kathmandu. Part of a glacier above Langtang village broke off and plummeted into the valley below. Witnesses report a powerful wind blasting snow, dust and building fragments over the village, suggesting that the avalanche was so powerful that it sent a pressure wave racing outwards, says Dorothea Stumm, a glaciologist at the International Centre for Integrated Mountain Development in Kathmandu. Satellite images reveal that a giant swathe of mountainside has been obliterated, right down to the river on the valley floor.

Elsewhere, Nepalese authorities worry that landslides might block rivers. This can form lakes that obstruct major roads, force people to evacuate their homes or even burst in a single catastrophic event. Last August, a landslide triggered by heavy rains dammed the Sun Kosi River in northern Nepal, killing roughly 150 people and causing widespread flooding.

At least four groups are using government and commercial satellite imagery to map such hazards. The recent quake caused landslides ▶

that blocked valleys, including several along the Trisuli River, which runs between Nepal and Tibet. "It looks to be quite risky there at the moment," says Nick Rosser, a landslide expert at Durham University, UK. "This will be the area of biggest impact when the monsoon starts, as rainfall totals there are among some of the highest in the country." In preparation, a team from the Chinese Academy of Sciences has been surveying landslide sites along roads that lead to Tibet. Another area of concern is a landslide-created lake on the Marshyangdi River, which runs above the Annapurna trekking circuit.

The monsoon in Nepal typically lasts from June to September, and fatal landslides happen mostly during that period. The amount of monsoon rainfall varies dramatically from year to year (D. N. Petley *et al. Nat. Hazards* **43,** 23–44; 2007). The fact that the earthquake struck in April may have been something of a saving grace, because dry soils are harder to dislodge than is wet ground. Many more landslides would have happened had the quake struck just a few months later, says Binod Tiwari, a geotechnical engineer at California State University in Fullerton.

The country was at high risk of landslides even before the tremor. Nepal rides atop the ongoing collision between India and Asia, a geological bust-up that pushes the Himalayas to ever-greater heights. The rugged terrain, unstable soils, heavy rains and mountain communities combine to make it one of the world's landslide hotspots.

The earthquake, now named the Gorkha quake, has worsened the situation. It ruptured the main Himalayan geological fault to the northwest of Kathmandu. By 11 May, more than 8,000 people had been confirmed dead — although that is many thousands fewer than experts had projected. Buildings in Kathmandu may have been sturdier than thought, or the ground may not have shaken quite as strongly as would be expected from the quake's magnitude.

Gentler shaking would also help to explain the relative paucity of landslides, says Marin Clark, a geomorphologist at the University of Michigan in Ann Arbor. "Either that, or the rocks are stronger than we estimated," she says.

But the risk of landslides remains high. After the magnitude-7.6 Chi-chi earthquake in Taiwan in 1999 and the magnitude-7.9 Sichuan earthquake in mainland China in 2008, the number of landslides soared for years as sediment continued to shift in fresh debris flows (B. Yu *et al. Eng. Geol.* **182,** 130–135; 2014). Authorities in Nepal need to prepare for the monsoon season by inspecting and monitoring the places most at risk, Rosser says. "That's where we need to be focusing," he says. "Places where we have landslides and there's a population." ■

POLITICS

What the UK election means for science

A Conservative majority, Scottish National Party rise and Liberal Democrat losses all have implications for research.

BY ELIZABETH GIBNEY

From an outright majority in Parliament for the Conservatives, to the decimation of the Liberal Democrats and the rise of the Scottish National Party, the UK general election on 7 May was full of surprises — many of which will have implications for science.

Scientists should expect an emphasis on austerity allied to economic growth — a pillar of the Conservative-led government of the past five years, which ruled in coalition with the Liberal Democrats. During that time, the science budget was frozen, and dropped in real terms. But there is support for science in Parliament, and an understanding of its relation to the economy, says Paul Nightingale, deputy director of the Science Policy Research Unit at the University of Sussex, UK. So cutting the research budget would be a "hard sell", he says. Instead, he expects "more explicit attempts to align research with economic growth".

Before the election, the Conservatives pledged to seek a "strong" deal at the United Nations climate negotiations in December that "keeps the goal of limiting global warming to two-degrees firmly in reach". The party also promised to end support for onshore wind farms and to encourage expansion of nuclear power and gas, including fracking.

67PHOTO/ALAMY

Former Liberal Democrat MP Julian Huppert was a vocal supporter of science.

Nick Hillman, director of the Higher Education Policy Institute in Oxford, UK, says that the reappointment of Theresa May as Home Secretary may trouble scientists. The last government's tough stance on immigration included cutting the post-study work visa for international students (see *Nature* **506,** 14–15; 2014). But he sees promise in the new science and universities minister, Jo Johnson, younger brother of the Mayor of London and new Member of Parliament (MP), Boris Johnson. Although not a scientist, Jo Johnson is pro-European and close to the Chancellor of the Exchequer, George Osborne.

A certain outcome of the Conservative win is a referendum by 2017 on whether to leave the European Union. Nightingale suspects that people will vote to stay in. If Britain did leave, it would probably not be cut out of European research programmes, says Kieron Flanagan, a science-policy researcher at Manchester Business School. However, it would feel the loss of cash from a different European pot, which Britain uses to fund science-related infrastructure.

Meanwhile, the Scottish National Party's increased representation from 6 to 56 seats may affect science across the United Kingdom. Growing Scottish nationalism is likely to further a focus on regional development that was part of the Conservative Party's manifesto, says Nightingale. This could bolster a trend to allocate science funding directly from the Treasury in London to regional projects, such as the UK National Graphene Institute in Manchester, he says, rather than through national funding agencies. However, David Price, vice-provost for research at University College London, warns that without an increase in the science budget, the regional agenda would be pointless.

Another defining moment was the crash in support for the Liberal Democrats, who lost 49 of their previous 57 seats, including some high-profile MPs. Keenly felt by many scientists was the loss of Julian Huppert, the former Liberal Democrat MP for Cambridge. Previously a biochemist at the University of Cambridge, he was vocal on science issues and popular with scientists. On 8 May, Huppert teasingly expressed fears about the future: "Scientists love control experiments — but I certainly didn't want a Tory government to show how effective the Liberal Democrats actually were in government." ■

ERIC MEOLA/CORBIS

Lightning is only the most visible product of clouds' intense electric fields.

PHYSICS

Rogue antimatter found in clouds

Aeroplane detects signature spike in thundercloud photons that does not fit any known source of antiparticles.

BY DAVIDE CASTELVECCHI

When Joseph Dwyer's aeroplane took a wrong turn into a thundercloud, the mistake paid off: the atmospheric physicist flew not only through a frightening storm but also into an unexpected — and mysterious — haze of antimatter.

Although powerful storms have been known to produce positrons — the antimatter versions of electrons — the antimatter observed by Dwyer and his team cannot be explained by any known processes, they say. "This was so strange that we sat on this observation for several years," says Dwyer, who is at the University of New Hampshire in Durham.

The flight took place six years ago, but the team is only now reporting the result (J. R. Dwyer *et al. J. Plasma Phys.*; in the press). "The observation is a puzzle," says Michael Briggs, a physicist at the NASA Marshall Space Flight Center in Huntsville, Alabama, who was not involved in the report.

A key feature of antimatter is that when a particle of it makes contact with its ordinary-matter counterpart, both are instantly transformed into other particles in a process known as annihilation. This makes antimatter exceedingly rare. However, it has long been known that positrons are produced by the decay of radioactive atoms and by astrophysical phenomena, such as cosmic rays plunging into the atmosphere from outer space. In the past decade, research by Dwyer and others has shown that storms also produce positrons, as well as highly energetic photons, or γ-rays.

"The insides of thunderstorms are like bizarre landscapes that we have barely begun to explore."

It was to study such atmospheric γ-rays that Dwyer, then at the Florida Institute of Technology in Melbourne, fitted a particle detector on a Gulfstream V, a type of jet plane typically used by business executives. On 21 August 2009, the pilots turned towards what looked, from its radar profile, to be the Georgia coast. "Instead, it was a line of thunderstorms — and we were flying right through it," Dwyer says. The plane rolled violently back and forth and plunged suddenly downwards. "I really thought I was going to die."

During those frightening minutes, the detector picked up three spikes in γ-rays at an energy of 511 kiloelectronvolts, the signature of a positron annihilating with an electron.

Each γ-ray spike lasted about one-fifth of a second, Dwyer and his collaborators say, and was accompanied by some γ-rays of slightly lower energy. The team concluded that those γ-rays had lost energy as a result of travelling some distance and calculated that a short-lived cloud of positrons, 1–2 kilometres across, had surrounded the aircraft. But working out what could have produced such a cloud has proved hard. "We tried for five years to model the production of the positrons," says Dwyer.

Electrons discharging from charged clouds accelerate to close to the speed of light, and can produce highly energetic γ-rays, which in turn can generate an electron–positron pair when they hit an atomic nucleus. But the team did not detect enough γ-rays with sufficient energy to do this.

Another possible explanation is that the positrons originated from cosmic rays, particles from outer space that collide with atoms in the upper atmosphere to produce short-lived showers of highly energetic particles, including γ-rays. "There's always like a light drizzle of positrons," says Dwyer. In principle, there could be some mechanism that steered the positrons towards the plane, he says. But the motion of positrons would have created other types of radiation, which the team did not see.

The team's data are a "cast-iron signature" of positrons, says Jasper Kirkby, a particle physicist who heads an experiment investigating a possible link between cosmic rays and cloud formation at the CERN particle-physics laboratory near Geneva, Switzerland. But "the interpretation needs to be nailed down". In particular, he says, the team's estimate of the size of the positron cloud is not convincing.

If Kirkby is right, and the cloud was smaller than Dwyer's team estimates, that could imply that the positrons were annihilating only in the immediate vicinity of the aircraft, or even on the craft itself. The wings could have become charged, producing extremely intense electric fields around them and initiating positron production, says Aleksandr Gurevich, an atmospheric physicist at the Lebedev Physical Institute in Moscow.

To answer these and other questions, Dwyer needs fresh observations of the innards of thunderclouds. To that end, he and others are sending balloons straight into the most violent storms, and the US National Science Foundation even plans to fly a particle detector on an A-10 'Warthog' — an armoured anti-tank plane that could withstand the extreme environment. "The insides of thunderstorms are like bizarre landscapes that we have barely begun to explore," says Dwyer. ■

GENOMICS

Microbiome privacy risk

The DNA of microorganisms living on a person's body could identify that individual.

BY EWEN CALLAWAY

Call it a 'gut print'. The collective DNA of the microbes that colonize a human body can uniquely identify someone, researchers have found, raising privacy issues.

The finding[1], published in *Proceedings of the National Academy of Sciences* on 11 May, suggests that it might be possible to identify a participant in an anonymous study of the body's microbial denizens — its microbiome — and to reveal details about that person's health, diet or ethnicity. A publicly available trove of microbiome DNA maintained by the US National Institutes of Health (NIH), meanwhile, already contains potentially identifiable human DNA, according to a study[2] published in *Genome Research* on 29 April.

The papers do not name individuals on the basis of their microbiomes — and predict that it would be difficult to do so currently — but they do suggest that those conducting microbiome research should take note.

"Right now, it's a little bit of a Wild West as far as microbiome data management goes," says Curtis Huttenhower, a computational biologist at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, who led the latest study[1]. "As the field develops, we need to make sure there's a realization that our microbiomes are highly unique."

Human-genomics researchers have grappled with privacy concerns for years. In 2013, scientists showed[3] that they could name five people who had taken part anonymously in the international 1,000 Genomes project, by cross-referencing their DNA with a genealogy database that also contained ages, locations and surnames.

In recent years, the microbiome's influence on our health and behaviour has become a hot research topic. Data from human-microbiome studies tend to end up in public repositories, but it was not clear whether microbiomes were permanent enough in individuals to identify them over time.

EYE OF SCIENCE/SPL

DNA from bacteria in human faeces could be used as a 'gut print' to identify individuals.

Working with publicly available data from the NIH Human Microbiome Project (HMP), Huttenhower's team searched samples taken from body sites, including the gut, mouth, skin and vagina, for combinations of microbial genetic markers that were both unique to a person and stable over time. (Although the HMP does not identify individuals by name, it is possible to compare a participant's first sample with a second one donated weeks or months later.)

Stool samples offered the best microbiome signatures; a person's first sample could be linked to their second sample 86% of the time. By contrast, skin samples could be accurately matched only about one in four times. The researchers note that DNA signatures based on individual strains of microbes did the best job of distinguishing people — much better than those based only on microbial species.

Still, Huttenhower concludes that it would be "exceptionally challenging to do anything with the microbiome data in a single study". The likeliest risk to privacy, he thinks, would come from a scenario in which someone had participated in two different microbiome studies that each contained different pieces of accompanying information, such as age and health status.

But microbiomes could also pose a privacy risk because they inevitably get jumbled up with human DNA. Although the NIH went to considerable lengths to weed human DNA out of its HMP database, a team led by computational biologist Jonathan Allen of Lawrence Livermore National Laboratory in California has found[2] that contamination is still rife. For example, the team found sequences known as short tandem repeats that tend to vary between individuals and are used for making DNA matches in forensics. It is not clear whether their presence in microbiome samples could constitute a precise DNA signature, Allen says, but the rise of publicly available DNA databases increases the likelihood. *Genome Research* agreed to publish the paper by Allen and his team only if the NIH removed the known human sequences from its database.

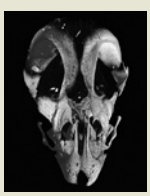
The odds of identifying someone on the basis of their microbiome is low, but researchers should take reasonable steps to protect privacy, says Yaniv Erlich, a computational geneticist at the New York Genome Center who led the team that identified[3] participants in the 1,000 Genomes study. Those who took part in the HMP were advised of the risk, says Amy McGuire, a bioethicist at Baylor College of Medicine in Houston, Texas. "I don't think there should be premature panic over this."

An overreaction could slow understanding of the microbiome. Laura Rodriguez, director of policy at the NIH's National Human Genome Research Institute in Bethesda, Maryland, says that as long as protections are in place, such as removing as much human DNA from the HMP as possible, "we would want to keep it in open access because of the value it adds to science". ■

1. Franzosa, E. A. *et al. Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1423854112 (2015).
2. Ames, S. K. *et al. Genome Res.* http://doi.org/4jt (2015).
3. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. *Science* **339,** 321–324 (2013).

MORE ONLINE

TOP STORY



Making chicken embryos with dinosaur-like snouts helps to explain evolution of beak go.nature.com/zj4wyt

MORE NEWS

- Why the polls got the UK elections wrong go.nature.com/djmijg
- Mobile-phone microscope detects eye parasite go.nature.com/fjykaw
- Five factors that will decide if Philae wakes go.nature.com/secttl

NATURE PODCAST

Latest from the Large Hadron Collider, and neurologist Oliver Sacks's new memoir nature.com/nature/podcast

BHART-ANJAN BHULLAR

ESPEN RASMUSSEN/PANOS

An Ebola survivor is helped out of a clinic in Sierra Leone built by Médecins Sans Frontières.

PUBLIC HEALTH

Ebola failures prompt WHO rethink

Health agency's annual meeting will address shortcomings in outbreak response highlighted by West Africa crisis.

BY ERIKA CHECK HAYDEN

As West Africa's Ebola epidemic wanes, a fever is building for reform of the World Health Organization (WHO). When the 194 member states of the United Nations body meet on 18–26 May in Geneva, Switzerland, for the 68th World Health Assembly, they will consider a number of proposals intended to make the WHO more nimble and effective at responding to fast-moving public-health crises.

At stake is the organization's future as the primary responder to global health emergencies. In September 2014, after the WHO's slow response in West Africa, the UN created a separate organization, the UN Mission for Ebola Emergency Response, to tackle the outbreak. And last month, UN secretary-general Ban Ki-moon appointed a panel on managing future health crises that is expected to issue a report by the end of this year. Observers note that previous WHO failures led to the formation of independent initiatives such as GAVI, the Vaccine Alliance, which vaccinates children in the developing world, and the Global Fund to Fight AIDS, Tuberculosis and Malaria.

"The worry would be that some other group will take contingency planning and emergency response away from WHO," says infectious-disease specialist Barry Bloom of the Harvard T. H. Chan School of Public Health in Boston, Massachusetts. "That has been the trend and leads back to the fundamental question of what is the role of WHO."

WHO director-general Margaret Chan has acknowledged that the agency made big mistakes in the Ebola epidemic. The outbreak revealed "inadequacies and shortcomings in this organization's administrative, managerial and technical infrastructures," she said in a January speech, which asked for reforms to the WHO and for an external review of the agency's performance. A draft of the review released on 8 May concluded that "WHO does not have the operational capacity or culture to deliver a full emergency public health response."

The main reforms up for discussion in Geneva include creating a US$100-million fund for response to fast-moving events such as the Ebola epidemic; setting up an international cadre of first responders to outbreaks; and setting guidelines for how aid groups, foundations, academic institutions and corporations can take part in WHO meetings. Such guidelines have been in the works for years as global public-health spending has increasingly shifted to non-governmental organizations. But the importance of private entities was brought to the fore by the Ebola outbreak, with groups such as Médecins Sans Frontières (also known as Doctors Without Borders) proving to be more effective than the WHO or governments.

The agency is also asking member states to boost its budget by 8% for 2016–17, after having received flat funding since 2012. And Chan wants to strengthen the International Health Regulations — rules agreed by member states in 2005 that require countries to set up basic outbreak-response mechanisms — but there are no specific proposals for how this would occur.

Many reformers attending the meeting are worried that even the enormity of the Ebola epidemic will not motivate the WHO membership to adopt reform, and are pessimistic about the effectiveness of such changes if they are approved. Some of the same ideas were proposed — but never adopted — after outbreaks of the respiratory syndrome SARS, H5N1 'bird flu' and H1N1 'swine flu' in the 2000s.

"I want to emphasize the importance of making these changes now, while the epidemic is fresh in our mind, and not wait, because the political momentum is with us now, and it will fade the same way it did with SARS and H1N1," says Lawrence Gostin, a health-law and policy specialist at Georgetown University in Washington DC who serves on a WHO committee charged with reforming the organization.

That sentiment is shared by Adam Kamradt-Scott, a health-security specialist at the University of Sydney in Australia. But he is pessimistic about the probability of change and frustrated about poor resources. "I doubt that we will see any significant reforms emerge," he says. "While governments complain on the one hand that the organization isn't effective, they also don't give the WHO sufficient resources or authority to do the job they've asked it to do."

The WHO relies on voluntary contributions for more than 70% of its budget; it is unlikely that countries would provide for the new task force or contingency fund at needed levels, says Kamradt-Scott.

Yet in the wake of the Ebola outbreak, governments and international organizations have signalled a willingness to provide funding to strengthen preparedness for health crises, says Josh Michaud, associate director of global health policy at the non-profit Kaiser Family Foundation in Washington DC. He notes that the United States last year approved more than $5 billion in emergency funding for Ebola response and that the World Bank has pledged to provide $650 million to help countries to rebuild after the epidemic. The United Kingdom has also said that it would provide $10 million to the WHO's proposed contingency fund.

"It's not easy or cheap to build up these capacities," Michaud says. "But there are signs that Ebola has sparked real change." ■

ANINDITO MUKHERJEE/REUTERS/CORBIS

Genetically modified mustard is being grown in fields in New Delhi.

BIOTECHNOLOGY

India eases stance on GM crop trials

States begin to permit field tests of transgenic plants.

BY SANJAY KUMAR

Five years ago, India was a hostile place for researchers testing genetically modified (GM) crops. Its government barred the commercial planting of a transgenic aubergine (a vegetable locally known as brinjal) after protests from anti-GM activists. Then it gave state governments the power to veto transgenic-crop field trials. The result: an effective moratorium on such trials. "We felt as if we had come up against a brick wall, and might as well chuck it in and do something else," says molecular biologist Bharat Char, who works for Maharashtra Hybrid Seeds Company (Mahyco), a firm in Jalna that pioneered the GM brinjal (and in which agricultural giant Monsanto holds a minority stake).

But under the government of Prime Minister Narendra Modi, voted into power a year ago, India has quietly changed course on GM field testing. In the past year, eight Indian states largely aligned with Modi's Bharatiya Janata Party have approved field trials of GM crops, between them allowing tests that include transgenic rice, cotton, maize (corn), mustard, brinjal and chickpea, according to documents seen by *Nature* (see 'GM crop trials'). "There is no better feeling than to know that your technology is performing in the field," says Char, who himself planted salt-tolerant GM rice in Maharashtra state in January.

DEVELOPMENT TENSIONS

The relaxed attitude to GM crop trials is not only reviving the enthusiasm of Indian biotech researchers — it will also be keenly watched around the world, says Dominic Glover, an agricultural socio-economist at the University of Sussex in Brighton, UK. "India's attitude towards transgenic crops has a symbolic importance beyond its borders," he says, because it epitomizes tensions that surround the use of GM technology in developing nations.

On the one hand, India must improve its agricultural productivity to feed its rapidly growing population. The country should thus embrace GM efforts to develop higher-yield crops that are resistant to pests or grow well in droughts or harsh environments such as salty soil, says biochemist Govindarajan Padmanaban, former director of the Indian Institute of Science in Bangalore.

On the other hand, India has more than 100 million farmers, who are concerned that if GM crops become prevalent, their livelihoods and the nation's food supply will increasingly rely on expensive, rapidly changing and proprietary seed technologies owned by large corporations, says Glenn Stone, an environmental anthropologist at Washington University in St. Louis, Missouri.

These tensions erupted in 2010, when farmers and anti-GM groups organized huge public protests that led to the brinjal ban (see *Nature* http://doi.org/bkt7dh; 2010). And they regularly flare up in criticisms of India's 2002 adoption of GM cotton, which contains genes to ward off certain insects. This is the country's only permitted commercial GM crop, but it is grown in such quantities that India is the world's fourth-biggest GM-crop producer, behind the United States, Brazil and Argentina.

SLOW PROGRESS

The new lenience on GM field trials has not reached all of India: more than 20 states and territories are still exercising their vetoes. In meetings between March and July last year, the Genetic Engineering Appraisal Committee (GEAC), part of India's environment ministry, granted permission to 80 field-trial applications, but state-government blocks meant that many of the trials were never begun.

These local bottlenecks have mainly hurt biotech researchers in India's universities and public-sector institutions, says Padmanaban: they retard the progress of domestic technology, whereas multinational firms such as Monsanto can test GM crops elsewhere. Akshay Pradhan, a geneticist working on transgenic mustard plants with the University of Delhi's Centre for Genetic Manipulation of Crop Plants, says that his team had its technology ready as far back as 2002 but is only now resuming field trials after a two-year hiatus. "This delay has surely thwarted its commercialization," he says. Agricultural scientists want faster progress: in February, some petitioned Modi to lift barriers including the requirement to seek state-government approval for field trials after getting the thumbs-up from the environment ministry.

But activists and non-governmental organizations (NGOs) argue that India should be wary of welcoming transgenic crops. They frequently raise concerns that such crops may be unsafe for the environment or human health, and that Indian regulators have conflicts of

SOURCE: GEAC DOCUMENTS SEEN BY *NATURE*

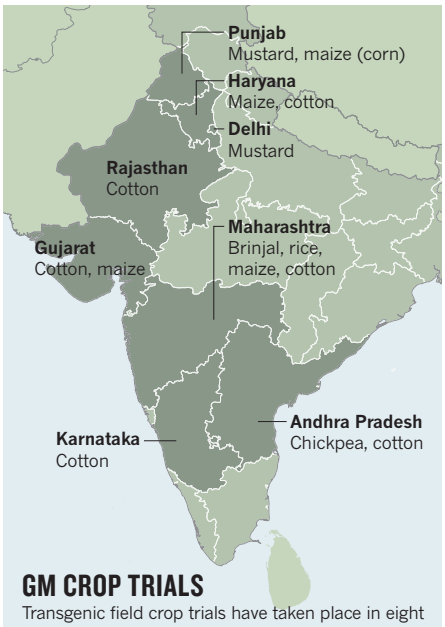
interest and have not put in place sufficient mechanisms to carefully monitor field trials. Similar criticisms were raised in 2012 by a technical committee convened by India's Supreme Court. The court is still considering a moratorium on planting GM crops (including in field trials), which anti-GM activists requested in a petition a decade ago.

One issue that critics and scientists agree on is the need for legislation to improve biotechnology regulations. A regulatory bill that failed to get through parliament in 2013 is now being revised, but could take two years or more to be passed, says Sunkeswari Raghavendra Rao, an adviser to the government's Department of Biotechnology.

COMMERCIAL CAUTION

India's government seems to be treading much more cautiously on commercial cultivation of transgenic crops than on field trials — although farmers in neighbouring Bangladesh began cultivating GM brinjal last year. (Some policy-makers who do not want to be named are concerned that the brinjal, or its seeds, will make its way into India anyway, over the porous Indo-Bangladesh border.)

But the government seems reluctant to engage in transparent debates about the pros and cons of pushing forward the use of GM biotechnology in India, as well as about the details of field trials being allowed in the country. Details of GEAC meetings that used to be publicly posted on its website now no longer appear online, and GEAC officials would not talk to *Nature* for this article. "I find this secrecy shocking and absurd," says Pushpa Bhargava, a retired molecular biologist whom the Supreme Court has nominated to attend GEAC meetings. And the government has come under fire for freezing the bank accounts of the Indian branch of environmental NGO Greenpeace. It has cited financial irregularities, which Greenpeace denies, but a widely leaked intelligence report prepared for Modi last year stated that the group's anti-GM campaigning was thwarting India's development. ■



GM CROP TRIALS

Transgenic field crop trials have taken place in eight Indian states and territories over the past year.

CORRECTIONS

The graphic in the News story 'Pluto mission hunts for hazards' (*Nature* **521,** 14–15; 2015) put Nix on the wrong orbital path. The correct image can be seen at go.nature.com/sbpxsu. The News story 'Pint-sized DNA sequencer impresses first users' (*Nature* **521,** 15–16; 2015) mistakenly stated that the MinION device has problems reading long, repetitive regions of DNA sequence. It should have said that those difficulties occur in sections of genome that are rich in long stretches of a single DNA base. And the reference list in the News story 'Mysterious galactic signal points LHC to dark matter (*Nature* **521,** 17–18; 2015) omitted two entries. The complete list can be seen at go.nature.com/mzopta.

SCIENCE IN INDIA

A special issue explores the enormous potential and major challenges for research in south Asia's superpower.

ANTON BALAZH/SHUTTERSTOCK

India is racing forward. With nearly 1.3 billion people and a steady growth rate, it is expected to become the world's most populous nation within a generation. Its gross domestic product more than tripled between 2000 and 2013, and its economy ranks third in the world in terms of purchasing power, behind only China and the United States. India's scientific production has also surged, with the number of published papers quadrupling over the same period.

But the country has far to go before it earns the status of a scientific superpower. By almost every metric — spending, number of researchers and quality of publications — India underperforms relative to developed nations and the ascendant economies to which it is most often compared, such as China and Brazil.

This week, *Nature* takes an unvarnished look at the challenges and opportunities for scientists in India. An infographic (page 142) assesses the country's strengths and weaknesses by comparing its research and development landscape with those of comparable countries. A News Feature (page 144) probes beneath the numbers, examining Indian successes in space, biotechnology and energy, as well as exploring bureaucracy, underfunding and other obstacles to higher education and scientific research.

Scientists have high hopes that Krishnaswamy VijayRaghavan, the new secretary of the Department of Biotechnology, can help to drive change in biomedical research. He is profiled on page 148. Ten Indian research leaders offer their suggestions for how to build their country's scientific capacity — from better funding, facilities, education and mentoring to fairer recruiting, more autonomy and a focus on local problems (page 151). Cheap and clean power will be key, say energy specialists Arunabha Ghosh and Karthik Ganesan (page 156). Only by tackling such basic issues can India hope to catch up with other rapidly advancing nations. ■

SCIENCE IN INDIA
A *Nature* special issue
nature.com/indiascience

INDIA
by the numbers

BY RICHARD VAN NOORDEN

Highs and lows in the country's research landscape.

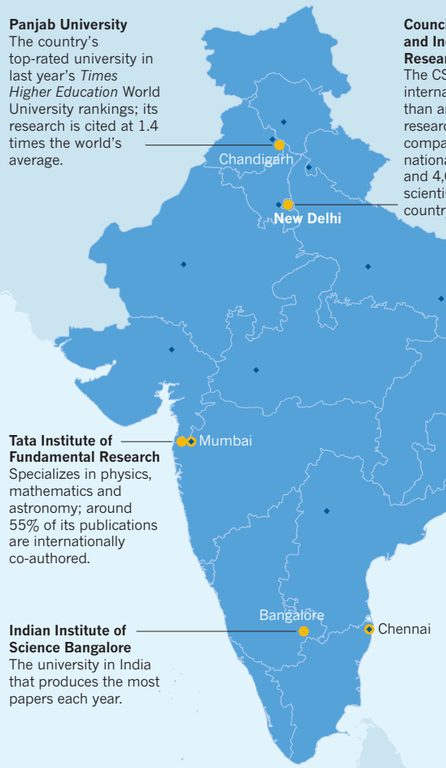
◆ Indian science is a study in contrasts. With its vast population and rapidly expanding economy, the country has ramped up scientific production at an impressive rate. India started the twenty-first century well behind Russia, France, Italy and Canada in terms of yearly publications and it now leads them all by healthy margins. It is quickly closing in on Japan.

Despite those gains, India is not yet a major player in world science. Its publications generate fewer citations on average than do those of other science-focused nations, including other emerging countries such as Brazil and China. Relative to its size, India has very few scientists; many Indian-born researchers leave for positions abroad and very few foreign scientists settle in India. The country invests a scant portion of its economy in research and development (R&D), and it produces relatively few patents per capita compared with other nations.

But there are bright spots. India boasts several world-class centres for science education, particularly the highly regarded Indian Institutes of Technology. Businesses in the country are investing more in R&D, which bodes well for future innovation. And more women are participating in science, although their numbers still fall far below those of men. ◆

Elite research centres

India's 700 or so universities vary tremendously in quality. To identify the leading science institutions, *Nature* looked at the citation rates in Elsevier's Scopus database for institutes that had produced more than 2,000 papers between 2010 and 2014.



Panjab University
The country's top-rated university in last year's *Times Higher Education* World University rankings; its research is cited at 1.4 times the world's average.

Council of Scientific and Industrial Research (CSIR)
The CSIR files more international patents than any other Indian research institute or company; it boasts 38 national laboratories and 4,600 active scientists around the country.

Indian Institute of Technology (IIT; ◆)
There are 16 IITs nationwide, accepting only elite students; acceptance rates are about 2%, compared with around 6% at Harvard University.

Indian Association for the Cultivation of Science
India's oldest research institute, founded in 1876, is where Nobel laureate C. V. Raman discovered the light-scattering effect that bears his name.

Tata Institute of Fundamental Research
Specializes in physics, mathematics and astronomy; around 55% of its publications are internationally co-authored.

Indian Institute of Science Bangalore
The university in India that produces the most papers each year.

Top 10 institutions

Number of papers in Scopus database 2010–14		Citation impact (World average = 1)
4,000	Panjab University, Chandigarh	1.4
4,400	Tata Institute of Fundamental Research, Mumbai	1.39
2,400	Indian Association for the Cultivation of Science, Kolkata	1.28
9,300	CSIR Chemistry & Physics, 5 locations	1.18
6,800	Indian Institute of Technology Bombay	1.15
10,800	Indian Institute of Science Bangalore	1.11
3,700	Indian Institute of Technology Guwahati	1.07
4,800	CSIR Industry & Standards, 12 locations	1.07
8,100	Indian Institute of Technology Kharagpur	1.06
6,700	Indian Institute of Technology Madras, Chennai	1.03

40%
of Indian-born researchers were working overseas in 2011 — the largest diaspora of any of the 16 countries in a survey of researchers (see *Nature* **490,** 326–329; 2012).

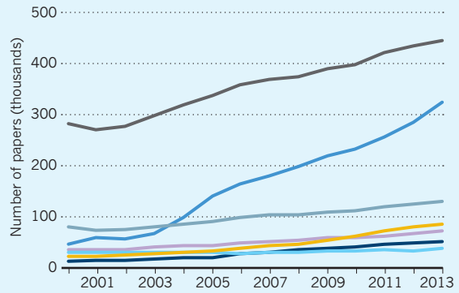
SCIENCE IN INDIA
A *Nature* special issue
nature.com/indiascience

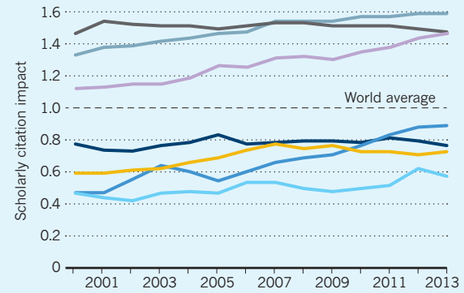
ELITE RESEARCH CENTRES: SCIVAL/SCOPUS; CSIR; IACS; *NATURE* **472,** 24–26 (2011); DESIGN: JASIEK KRZYSZTOFIAK/*NATURE*

Publications

Since 2000, India has almost quadrupled its scholarly output, but that rate is surpassed by Brazil's and China's. India underperforms relative to its gross domestic product (GDP) and population. And its scholarly impact remains low: in 2013, it was nearly 30% below the world's average.

US China UK India Italy Brazil Russia



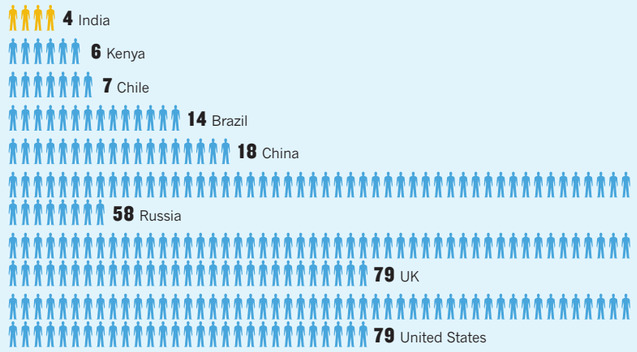


PUBLICATIONS: SCIVAL/SCOPUS; SCHOLARLY CITATION IMPACT: SCIVAL/SCOPUS; WORKFORCE: UNESCO; GRANTS TO WOMEN: DST

Workforce

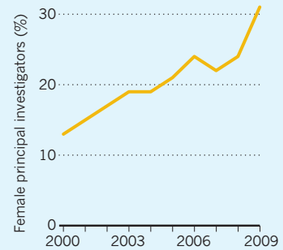
With only 200,000 full-time researchers (14% of them female) in a population of nearly 1.3 billion, India ranks below Chile, Kenya, and many other countries in terms of the density of its scientific workforce.

One researcher per 10,000 labour force



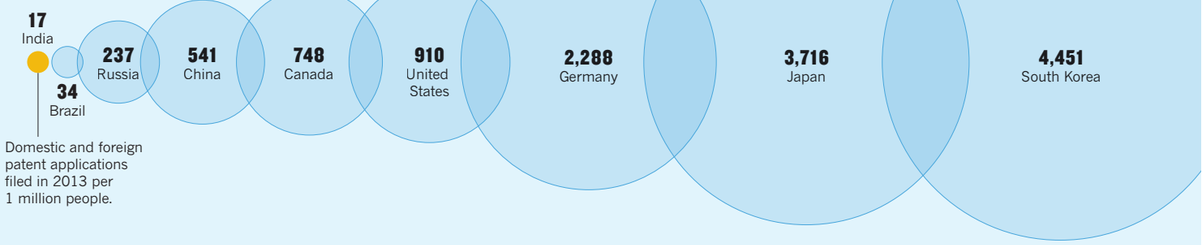
Grants to women

More women are winning funding from competitive government grant schemes, according to India's Department of Science and Technology.



Patents

India is one of the world's leading filers of patents but it registers far fewer applications per capita than any other top-filing nation. Multinational firms in India have boosted the country's rate of filing.



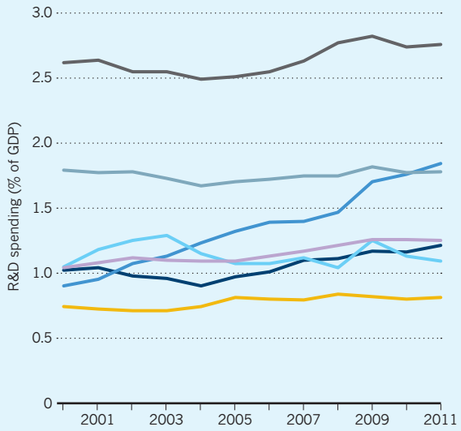
Domestic and foreign patent applications filed in 2013 per 1 million people.

PATENTS: WIPO; R&D INVESTMENT: UNESCO/DST; SPENDING PER RESEARCHER: UNESCO

R&D investment

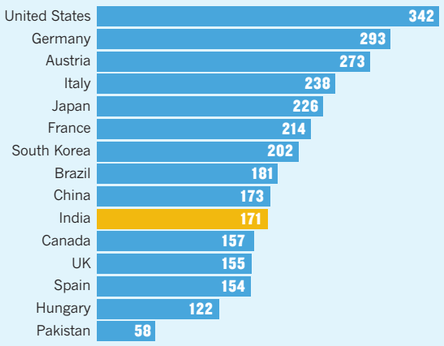
Whereas China's research spending has shot up to almost 2% of its GDP, India's languishes at around 0.9%, a figure that has changed little in more than a decade and lags behind both Brazil's and Russia's.

US China UK India Italy Brazil Russia



Spending per researcher

India spends much the same per researcher as many other countries; figures are normalized for purchasing power and are in thousands of US dollars.



INDIA'S SCIENCE TEST

The south Asian superpower has made great strides in research and development, but it has a long way to go.

BY T. V. PADMA

With her jeans, T-shirt and spirited attitude, Tapasya Srivastava could pass for a student as she works in her brightly lit cancer-biology lab on the University of Delhi South Campus. Srivastava, who oversees a team of eight researchers, is thrilled that she earned "a small research space of my own" in 2010, while still in her thirties. "With a decent list of publications under my belt, I am one of the few who have studied and undergone training entirely in India," she says.

Eight kilometres away, in the chemical-engineering department of the Indian Institute of Technology Delhi, Shalini Gupta's team is developing sensors to detect early-stage sepsis and typhoid. Gupta did her doctorate in the United States but returned to India to focus on its needs: "I am more connected to society and its challenges," she says.

Srivastava and Gupta are part of a wave of young Indian scientists convinced that they can do high-quality research at home rather than having to move abroad. Such optimism reaches all the way to the top: in January, Indian Prime Minister Narendra Modi told an assembly of scientists to "dream, imagine and explore. You will have no better supporter than me."

India has much to be proud of. Last year, it became the first to reach Mars on its initial attempt. It boasts a thriving pharmaceutical industry that produces low-cost medications that are desperately needed by the developing world. And in his first year in office, Modi launched an ambitious plan to make India a leader in solar power.

Such successes cannot hide the huge challenges facing this country of 1.3 billion people, which leads the world in tuberculosis incidence and maternal deaths, and lacks electricity for one-quarter of its citizens. India is expected to become the world's most populous nation within a generation, and it will require a robust science and technology sector to supply the needed energy, food, health care, jobs and growth. Yet researchers in India and abroad say that the country has a relatively weak foundation in science and engineering.

Indian research is hampered by stifling bureaucracy, poor-quality education at most universities and insufficient funding. Successive governments have pledged to increase support for research and development to 2% of India's gross domestic product (GDP), but it has remained static at less than 0.9% of GDP since 2005. Despite its huge size, India has a relatively tiny number of researchers, and many of its budding scientists leave for other countries, never to return. Only by tackling its systemic problems can India compete with other emerging powerhouses such as Brazil and China.

"The density of scientists and engineers in India is one of the lowest in the world," says Sunil Mani, an economist at the Centre for Development Studies in Trivandrum, who is assessing Indian science and engineering for an upcoming report by the United Nations Educational, Scientific and Cultural Organization. "There are very many important areas where we are not able to do research."

SCIENCE IN INDIA
A *Nature* special issue
nature.com/indiascience

SPACE TO GROW

In one of the cleanest rooms in India, Mylswamy Annadurai is busy conducting fitness tests on a 750-kilogram patient — a gleaming satellite called ASTROSAT. The probe is strapped to a table, where it is being shaken at six times the strength of gravity to simulate the intense forces of lift-off. ASTROSAT must also pass tests in extreme high and low temperatures and vacuum conditions, followed by checks on its solar arrays and antennas. If all goes well, the satellite will blast into orbit by September, armed with two telescopes and four other instruments to study both nearby and distant stars.

Annadurai, who is head of the satellite centre of the Indian Space Research Organisation (ISRO) in Bangalore, says that ASTROSAT will

India is one of the leading nations in wind power and it has ambitious goals for increasing solar power over the next decade.

PRISMA BILDAGENTUR AG/ALAMY

be India's "first full-fledged science mission" in space. It will carry instruments ten times heavier than those on India's first mission to the Moon, 2008's Chandrayaan-1, and its 2014 Mars Orbiter Mission, nicknamed Mangalyaan.

With its run of recent accomplishments, India has earned international acclaim for its ambitious space programme, which includes launch vehicles, communication satellites and one of the world's largest constellation of remote sensing satellites, as well as its science missions. Since ISRO was founded in 1969, the government has invested heavily in it, and even established a dedicated university in 2007 to train personnel. "The ISRO technical test, assembly and launch facilities are first class," says Paul Spudis, senior staff scientist at the Lunar and Planetary Institute in Houston, Texas, who was the principal investigator for one of Chandrayaan-1's experiments.

Chandrayaan-1 carried an orbiter and a 35-kilogram probe that took images as it smashed into the Moon at high speed. ISRO plans to follow it in 2017 with Chandrayaan-2, which will gently set down a lander and a six-wheeled rover; together with an orbiter, they will study the composition of the Moon's surface. Up next after that is the Aditya mission to study the Sun's corona, in 2018.

Spudis is critical of last year's Mars mission, calling it "largely irrelevant" and saying that it would have been better to return quickly to the moon. ISRO, he says, "seems to lack a strategic vision of what it wants to accomplish in space". But the agency counters that it is pursuing several missions in parallel; the Mars mission just proceeded faster than Chandrayaan-2.

And the success in reaching Mars has convinced others at ISRO that they can carry out world-class space-science missions, says Annadurai. "The Mars mission experience has once again strengthened our belief that we can."

BIOTECH BONANZA

In Genome Valley, a biotechnology park in Hyderabad, entrepreneur Krishna Ella is confounding expectations. Ella returned home from the United States in 1996 with a 12-metre shipping container filled with vaccine-making equipment to support his grand plan of producing a US$1 vaccine for hepatitis B. That goal, which would make his vaccine at least an order of magnitude cheaper than the available one, struck investors as crazy, he says. But within three years, Ella's company Bharat Biotech International Limited (BBIL) succeeded in producing the Revac-B+ hepatitis vaccine at $3 a dose, which has since dropped to 30 cents per dose. It went on to develop vaccines against Japanese encephalitis, rabies, haemophilus influenza virus B and, most recently, rotavirus. Each costs barely a dollar per dose.

Affordable medicines are the cornerstone of India's health-care sector, where publicly funded hospitals struggle to provide treatment. The country has long battled infectious diseases such as tuberculosis, malaria and dengue, but is now facing rising numbers of non-communicable illnesses, including diabetes and coronary heart disease. A 2014 report from the World Economic Forum and Harvard School of Public Health estimates that non-communicable diseases and mental illness could cost India $4.58 trillion by 2030.

Low-price vaccines and generic drugs have helped India to carve out a niche in the international pharmaceutical industry. The global medical charity Médecins Sans Frontières (also known as Doctors Without Borders), which relies on Indian generics for 80% of its anti-HIV drugs, hails the country as the "pharmacy of the developing world". Other international organizations, including the UN children's charity UNICEF and the Global Fund to Fight AIDs, Tuberculosis and Malaria, routinely use Indian vaccines and generic drugs to treat infectious diseases (see *Nature* **468,** 143; 2010).

But India is battling criticism over the quality of some of its pharmaceuticals. In 2012, for example, the World Health Organization took BBIL's hepatitis B vaccine and oral polio vaccine off the list of drugs preapproved for use by the UN. Ella says that the issues related to documentation submission and that they have since been sorted out. The vaccines are now back on the list.

In 2014, the US Food and Drug Administration (FDA) sent warning letters to seven Indian firms over various concerns relating to pharmaceutical production there. An FDA spokesperson told *Nature*: "While some Indian companies meet US product quality standards, others have been found to lack sufficient controls and systems to assure drug quality, both of finished product and active ingredients." The FDA has an India office to work closely with Indian drug regulators to solve those problems.

And some in the biotech sector warn that India has a long way to go to create a thriving enterprise in developing new drugs. The country's success in the generics industry relies on a different set of skills: reverse-engineering pharmaceuticals created elsewhere by breaking them into their components and remaking them through cheaper routes.

"The challenge for the sector will be to graduate from reverse engineering to new-drug discovery," says Pallu Reddanna, a biotechnologist at the University of Hyderabad. "There is need for incentives and promotion of academy–industry interactions."

The government and private sector are trying to jump-start such efforts by setting up incubators that help transfer university and lab know-how to industry, and provide infrastructure and financial support to start-ups. Such incubators are the "greatest changer in the drug-discovery sector in India", says P. Yogeeswari, a chemist at the Hyderabad campus of Birla Institute of Technology and Science.

Krishnaswamy VijayRaghavan, secretary of the government's Department of Biotechnology, commends "incredible growth" in India's biotech entrepreneurship — despite the lack of big drug companies and the relatively low domestic investment in drug discovery. International and industry collaborations with academia are helping to advance the sector, he says (see page 148).

In 2013, the department started two major projects seeking drugs for drug-resistant tuberculosis and chronic disorders such as heart disease. In early leads, scientists have zeroed in on some human proteins that are crucial for the survival of multidrug-resistant tuberculosis strains. Proof-of-concept studies in mice have demonstrated that targeting such host proteins could help to kill the drug-resistant strains, says VijayRaghavan (S. Jayaswal *et al. PLoS Pathog.* **6,** e1000839; 2010). "We are at an exciting early applied stage," he says.

POWER HUNGRY

Nearly 2,000 kilometres north of Genome Valley, 9.7 hectares of solar panels cover a building in Punjab state, generating 7.5 megawatts of electricity. This project is India's largest roof-top solar installation that is connected to an electrical power grid, and it signals India's outsize ambitions in renewable energy.

Coal supplies two-thirds of the electricity in India and will remain king for some time. But the government has set aggressive goals for installing solar-energy capacity. In 2014, Modi's government announced that it would develop 100 gigawatts of solar-energy capacity by 2022. This is a huge leap from the existing 3.7 gigawatts of solar capacity — just 1.4% of India's total electricity generation today.

> "WE ARE CAUGHT IN A VICIOUS CIRCLE OF MEDIOCRITY."

"India is one of the most attractive markets in the world," says Pashupathy Gopalan, Asia Pacific head of SunEdison, a global solar-energy company based in Maryland Heights, Missouri, which is joining Adani Enterprises of Ahmedabad to build India's largest solar-panel-manufacturing facility. "We are entering a new era where solar electricity is competitive and has achieved 'socket parity' with other sources of energy in India."

There are other big international collaborations. The Solar Energy Research Institute for India and the United States was established in 2012 to target emerging research areas. In one project, researchers are trying to generate solar thermal power by using sunlight to heat up a highly compressed fluid form of carbon dioxide so that it turns electricity-generating turbines. This could be used in much smaller plants than conventional steam-driven turbines.

But some analysts say that India suffers from "gigawatt obsession". The focus on giant solar plants comes at the expense of smaller facilities that do not require large parcels of land, but could provide electricity to isolated towns, even without being connected to the grid.

"The gigawatt rush must pay attention to the pace with which the capacity is to be built in India," says Satish Agnihotri, former secretary of India's Ministry of New and Renewable Energy. Plans to build large plants could run into opposition in densely populated or heavily farmed areas, and in remote areas it can be difficult to hook gigawatt projects up to the electrical system.

News and debates about the government's current focus on solar power have overshadowed past successes in wind energy. India has more than 23 gigawatts of installed wind-power capacity, which puts it roughly even with Spain as the world's fourth biggest producer. And Mumbai-based Suzlon is the world's seventh-largest turbine manufacturer.

India has been able to develop its wind power in part because of long-term government policies and financial incentives, as well as a growing interest from independent power producers and financiers, says Shantanu Jaiswal, lead analyst at Bloomberg New Energy Finance in New Delhi. But some of the concerns about solar power also hamper wind projects, which face difficulty acquiring land, encounter lengthy permitting processes and often have trouble connecting to the electrical power grid.

EDUCATION OUTLOOK

Back on her leafy campus in Delhi, Srivastava and her fellow young faculty members are less concerned about big national projects than about producing their own high-quality research. They are lucky, they acknowledge, to work in one of India's top federally funded universities, which has superior faculty members and equipment.

Others are not so fortunate. India has some 700 universities of varying quality, from the elite institutions funded by the central government to more than 300 state universities and about 200 private ones. "The landscape of science education is uneven," says Sri Krishna Joshi, former director-general of India's Council of Scientific and Industrial Research (CSIR) and former chair of the advisory committee of the University Grants Commission, which funds and oversees university education in India.

In the top institutions, he says, "science students are doing world-class research, publishing in leading journals and boosting the global reputation of our country". National scientific research institutes and leading universities have all contributed to India's growing strengths in research: the country's output of scientific publications quadrupled between 2000 and 2013.

Even so, India is not keeping pace with some other emerging nations, which have increased their scientific output more quickly (see page 142). And the advances in India's global science metrics mask some signs of declining quality in university science education, especially at the cash-starved universities funded by state governments that account for the majority of India's science undergraduates, says Joshi. Publicly supported universities depend on the Ministry of Human Resource Development for funds, and the higher-education budget was hit by a 3% cut in the 2014–15 budget cycle.

"Lack of even bare, minimal and sustainable funds for teaching, let alone research, has seriously plagued the quality and standards of science education," says Krishna Ganesh, a chemist and director of the Indian Institute of Science Education and Research in Pune, one of five top universities set up in India since 2006.

Many students at state universities are receiving a substandard education, says Joshi. "Here, there are no good science teachers, no good Indian textbooks, and most of the science laboratories are poorly equipped," he says.

Graduates celebrate at the University of Delhi, a top institution. The majority of science students in India graduate from lower-quality universities that lack funding.

RAJ K. RAJ/*HINDUSTAN TIMES* VIA GETTY

"We are caught in a vicious circle of mediocrity," says geneticist Deepak Pental, former vice-chancellor of the University of Delhi.

Most analysts are concerned over the plight of science departments in state universities. At the University of Calcutta, for example, even procuring a laptop involves endless red tape, says physicist Amitava Raychaudhuri. At some other institutions, support from funding agencies helps to purchase equipment, but there is a shortage of qualified faculty members to train the students.

Beyond questions of quality, the quantity of available university spots is a persistent problem. India has gone through a university building boom, but there still is a huge shortage of slots for students (see *Nature* **472,** 24–26; 2011).

"There is a rise in the number of students going for higher education in India, which reflects the rising aspirations of its society. But this rise should be matched by better infrastructure and financial support," says Joshi.

RESEARCH INVESTMENTS

Investments in science have also dragged. India's research intensity — the share of its gross domestic product devoted to research and development (GERD) — remains lower than those of many other nations, including Brazil and Russia. Twenty years ago, India's GERD exceeded China's. Now, it is less than half.

But those numbers do not tell the whole story, says Ashutosh Sharma, secretary of the government's Department of Science and Technology — one of India's largest research-funding agencies. "The total funding is, perhaps, not as poor as it seems in terms of absolute numbers, because the number of full-time scientists doing research is also low."

India averages about 4 full-time researchers per 10,000 people in the labour force, whereas China boasts 18 and nations with advanced science and technology sectors have around 80. "India spends about $150,000 per scientist per year, which is probably not too far from the optimal levels," says Sharma.

India's notorious bureaucracy deserves part of the blame for the problems afflicting science education and research. The administrators of several state universities are political appointees rather than leading academics. "Often the appointed person has never been exposed to a good university in India or abroad," says Kizhakeyil Sebastian, chair of the science-education panel of the Indian Academy of Sciences in Bangalore.

"There is over-bureaucratization within the universities and their controlling bodies," says Pental. It often takes two years to recruit an academic after announcing an open post, which means that the best applicants can slip away, says Raychaudhuri.

The governmental quagmire has begun to affect some elite national research institutes, too. Of the 38 national laboratories that are part of the CSIR, only 25 have full-time directors. The rest are making do with acting directors, or temporary arrangements.

Even the CSIR headquarters in New Delhi has been without a full-time leader since January 2014. Interim director-general Madhukar Garg says that "the current situation is indeed challenging. CSIR is the backbone of scientific and technological research in the country. In case the prevailing scenario continues, it will affect the national innovation system as a whole."

Sharma acknowledges that red tape is "all-pervasive", but he says that the challenges are not bogging down Indian science. "In terms of output indicators such as the number of papers per dollar spent, Indian science is among the very top performers in the world," he says.

And there are some signs that India might be slowing its debilitating brain drain. Although the vast majority of Indians who obtain science doctorates in the United States remain there for at least 5 years after graduation, the proportion has declined: from 89% in 2001 to 82% in 2011, the most recent year for which data are available.

Kaustuv Datta, a geneticist at Delhi University South Campus, is one of those who returned. Datta may "hate the red-tapism" at universities in India, but he still prefers doing research back home. "My parents are here, in India. And academics have a strong, positive influence on the next generation of students," says Datta. "I wanted to make that contribution in India". ■

T. V. Padma *is a science journalist in New Delhi.*

THE ANTI-BUREAUCRAT

K. VIJAYRAGHAVAN IS DETERMINED TO CUT THROUGH RED TAPE AND BUILD UP BIOLOGICAL SCIENCE IN INDIA.

BY APOORVA MANDAVILLI

On 12 April, Krishnaswamy VijayRaghavan posted an update to his more than 2,500 Facebook friends. It announced a bold plan from India's Department of Biotechnology (DBT) — the agency that VijayRaghavan leads, and the country's largest funder of biomedical research — to establish a new marine-biology institute and research stations along India's vast coastline. Within hours, 500 people had 'liked' the post and more than 60 had left comments of congratulations.

Only one offered a critical note. A graduate student said that starting programmes is all well and good, but the DBT must hold the researchers whom it already funds accountable for the quality of their science. Shortly after, VijayRaghavan replied: "Your words are very wise and correct! Thank you. We must keep your points in mind if we are to get maximum for our Rupee and have quality science."

It is rare for a public official to be so responsive and open to criticism, especially in a country as steeped in bureaucratic hierarchy as India, says biologist Inder Verma at the Salk Institute for Biological Sciences in La Jolla, California, who has served as a scientific adviser to the Indian government since the 1980s. Yet almost anyone who contacts VijayRaghavan by Facebook, Twitter or e-mail gets a personal response in minutes. "Vijay is a breath of fresh air," Verma says.

VijayRaghavan is more than that. He is a respected fly geneticist and administrator who helped to build the National Centre for

SAM MOHAN

Biological Sciences (NCBS) in Bangalore, one of India's most prestigious institutions, from the ground up. In January 2013, he left his job as NCBS director and moved to New Delhi to lead the DBT. He says that he wants to inject rigour into Indian science and train scientists to work together on tractable problems. As grand visions go, his can seem muted, almost modest. "I'm not going to be stupid and try something completely nutty; I'm going to try something within my grasp," he says.

Researchers are optimistic about what he might be able to achieve. "It's very rare to have a scientist of Vijay's calibre heading a government department," says Jyotsna Dhawan, a stem-cell biologist who worked with VijayRaghavan for seven years. "So I think all of us in the scientific community have very high hopes."

But they also recognize the challenges, which include wrangling with New Delhi's murky politics — known for ensnaring plans in red tape — and the DBT's long, painful grant-review process. In the past couple of years, the Ministry of Finance has made it difficult for the agency to honour even approved grants. And although the DBT is a major funder of extramural research, the money that it actually gets each year — a little more than 14 billion rupees (about US$225 million) — is a fraction of that commanded by analogous agencies elsewhere, such as the US National Institutes of Health.

Given the challenges, even the most ardent well-wishers are holding their applause. "It's not entirely apparent to me what an individual, even one so dynamic and forward-looking as VijayRaghavan, can do to cut through the red tape," says Dhawan.

A PASSION TO LEARN

A self-described "air-force brat", VijayRaghavan grew up all over India, moving every few years. He was hungry for knowledge, and, as a teenager, used to cycle to his local branch of the British Council or the US Information Service — the main sources of foreign publications in those days — and read every book and magazine that he could find. "In the pre-Internet days, that was my food," he says.

After studying chemical engineering at the Indian Institute of Technology Kanpur, VijayRaghavan was preparing to leave for a bioengineering PhD programme in Switzerland when he chanced on an article by renowned molecular biologist Obaid Siddiqi on using genetics to understand the nervous system. It was a departure from the work that VijayRaghavan had originally planned to do, but, he says, "I found the formalism of genetics easy to grasp, and that excited me very much."

He sought out Siddiqi at the Tata Institute of Fundamental Research (TIFR) in Mumbai, where he began a PhD programme. It was, at the time, a place that afforded considerable freedom to its students. "You did what you pleased and you joined whomever you wanted to for your research," VijayRaghavan says. "It was an exhilarating experience."

But there was a growing air of complacency and nepotism at the TIFR that frustrated Siddiqi. For years, he had been planning to build a new institute, and he saw a natural ally in VijayRaghavan. The pair began to hatch plans, even as VijayRaghavan embarked on further training at the Medical Research Council Laboratory of Molecular Biology in Cambridge, UK, and later, undertook a post-doctoral fellowship at the California Institute of Technology (Caltech) in Pasadena.

> "I THINK ALL OF US IN THE SCIENTIFIC COMMUNITY HAVE VERY HIGH HOPES."

Elliot Meyerowitz, VijayRaghavan's adviser at Caltech, says that lab members routinely tried to flummox foreign postdocs with US slang and customs, but they could never rattle VijayRaghavan. "I don't know whether he understood, or if he was just so cool, we didn't know he didn't understand," Meyerowitz says. VijayRaghavan says that he did understand, thanks to his time devouring British and US magazines.

FROM THE GROUND UP

In 1988, VijayRaghavan left Caltech and returned to India to head a lab at the TIFR, and he, Siddiqi and a handful of other scientists laid groundwork for the research centre in Bangalore. They wanted the institute, which would be named the NCBS, to be different from any in India before it. Siddiqi became founding director, but VijayRaghavan and a few others were closely involved in its development. "We were in the trenches together — young, some very talented, all very driven," VijayRaghavan says. "We had a sense of rebellion."

From the start, VijayRagahvan wanted to recruit people trained in multiple disciplines who were focused on cutting-edge techniques, such as single-cell analysis, says statistician Partha Majumder. "This trait of being able to look way into the future is what sets him apart."

In 1991, VijayRaghavan moved to Bangalore to launch the NCBS's first lab. Over the next year, two more faculty members joined him. The entire centre was a "shack", he recalls, situated on one floor of the radio-astronomy building at the Indian Institute of Science. VijayRaghavan had to cycle 1.5 kilometres to the nearest biology lab to photograph his DNA gels. "We had an absolute ball of a time," he says.

Along with building the institute, VijayRaghavan was strengthening his scientific reputation. He borrowed equipment to set up a series of elegant genetic experiments that would enable him to write several high-profile papers defining specific events in *Drosophila* muscle development.

Other faculty members, such as Gaiti Hasan, M. K. Mathew and Jayant Udgaonkar, were publishing groundbreaking papers in cell signalling and protein folding, which in turn helped to entice other scientists to join the NCBS. "We made extraordinarily rash promises that we would do everything for them, which we did," VijayRaghavan says.

In 1993, for example, VijayRaghavan learned that Satyajit Mayor, a cell biologist in New York City whom he was trying to recruit, needed a pricey Zeiss inverted microscope. VijayRaghavan had been promised some equipment for his own lab through a grant from the Rockefeller Foundation in New York City, but stringent rules from the TIFR and the Indian government had held up the procurement for years. He changed his request to get the microscope instead.

All that Mayor knows of the negotiation is that he sent an e-mail to VijayRaghavan one night telling him that he would not be able to join the institute without that type of microscope. He woke up the next morning to VijayRaghavan's reply: "It'll be here when you arrive." Mayor joined the NCBS about 18 months later.

UNITED BY SCIENCE

VijayRaghavan took over from Siddiqi as director of the NCBS in 1996. As the institute grew, he strove to build a democratic system, in which even graduate students had a say, and criticism was not just accepted, but expected. Before moving ahead with any new plans, he always made sure that he "brought people along with him", says Mayor, now the institute's director.

In 1999, for example, the leaders of the institute were considering adding a master's programme in wildlife ecology and conservation. At first, only 3 or 4 of its then 22 faculty members were in favour of the idea. At a meeting, VijayRaghavan carefully listened to the pros and cons, and was ultimately able to convince everyone, recalls Mayor. "Everybody left that meeting feeling like we'd done the right thing," he says. "The way the discussion and the dialogue and the arguments were put across, quite masterfully, was so Vijay."

The programme has become one of the institute's most successful, with eight field stations across the country and faculty members drawn from the United States and Germany. By the time VijayRaghavan left, the NCBS was widely regarded as one of India's

COURTESY OF K. VIJAYRAGHAVAN AND TIFR

K. VijayRaghavan (bottom right) in a class photo at the Tata Institute of Fundamental Research in Mumbai in 1980. Obaid Siddiqi sits top left.

leading research organizations.

But the NCBS, and a few other select institutes, are exceptions in India. Much of the country's science is beset with the same problems it has had for decades — interminable waits for reagents, or a granting scheme that places a 3-year limit on funding, forcing researchers to write new applications in unrealistic cycles.

VijayRaghavan's predecessor at the DBT, Maharaj Kishan Bhan, had done much to modernize Indian research and make it more independent, including helping to develop a low-cost rotavirus vaccine and setting up an organization to support entrepreneurs. "If Bhan was not able to succeed in some places," says Verma, it was because of limited resources and not having enough people to support his vision. "Perhaps he bit off more than he could chew — the same could happen to Vijay."

BACK TO BASICS

Since his arrival at the DBT, VijayRaghavan has unveiled a few plans. Financially, Indian science is no match for that in the United States, Europe or China, something that he freely acknowledges. But he says that India can make big gains by capitalizing on its advantages and by collaborating with others.

His main priorities are to invest in basic research areas — such as computational biology, in which India is already strong — to break down the barriers between disciplines and to improve training for all scientists.

"I think he's got a very, very strong vision about the importance of fundamental and basic science," says Eve Marder, a neuroscientist at Brandeis University in Waltham, Massachusetts, who serves alongside VijayRaghavan on the scientific advisory board for the Janelia Research Campus in Ashburn, Virginia.

The DBT's marine-biology initiative exemplifies this vision. The effort, intended to chart biodiversity and identify compounds for biotechnology development, is the DBT's — and VijayRaghavan's — brainchild. But it involves the Indian Space Research Organisation in Bangalore and the Ministry of Earth Sciences in New Delhi, both first-time partners for the DBT. In addition, the French National Centre for Scientific Research and the Pierre and Marie Curie University in Paris will help to train Indian researchers. The project is part of VijayRaghavan's strategy to compensate for the DBT's limited budget by partnering with every ministry that has funds allotted for science and technology, including the sanitation, maternal-health and nutrition ministries.

In the next year, he plans to roll out an Indian body modelled after the European Molecular Biology Organization — a scientific society that would promote India as a hub for international collaborations and offer online training for scientists at all levels. He is also encouraging local collaboration on training and research. In the Delhi area, for example, he plans to persuade the leaders of well-established immunology, mathematics, engineering and medical institutes to work together. "Boom!" he says. "Within a few years, you're going to have extraordinary-quality people being trained, both engineers and clinicians".

In the short term, he wants to play to India's strengths. Getting India's thriving community of mathematicians and computer scientists to work on problems in biology, for example, could help the country to gain an edge in bioinformatics and quantitative biology — fields that do not typically require as much funding as bench biology.

This is all easier said than done, Vijay-Raghavan admits, but he intends to use financial incentives and disincentives — what he calls "fire in the belly" and "fire in the rear" — to make it happen.

In the past, the DBT has set out strict budgetary allocations at the beginning of each year, and had little flexibility in later months. But last year, VijayRaghavan set aside a pot of about $33 million from the DBT's annual budget. The money can be used to respond to innovative ideas, reviewed by international experts. This is sure to aggravate some institutes that are used to being well-funded, but VijayRaghavan is unmoved. "It's about time that we recognize excellence and recognize shoddiness," he says.

VijayRaghavan is also tackling the grant-review process. He is streamlining the DBT's online application system, he has created timelines for submitting grants and has introduced the DBT's first open-access and conflict-of-interest policies.

"When people complain about problems in India, it's rather astounding how few of us are actually doing something about them," he says. "If you actually start doing something about anything, the situation changes."

But some colleagues have concerns. Mayor says that VijayRaghavan's desire for consensus and harmony could prove a weakness. "He is so keen to be extremely positive about everyone," Mayor says. "When you're operating in the real world where you have to get things done, that, I would say, is a bit of a problem."

And in a government department, VijayRaghavan will not be able to hand-pick people who share his vision, as he did at the NCBS. "I'm a little concerned that if he doesn't have that, he will burn out, because he will try to do it all himself," Mayor says.

VijayRaghavan shows no sign of burnout yet. He still maintains a lab, albeit a lean one, conferring with his team in the evenings through Skype and returning to Bangalore every weekend. Besides the tweets and Facebook posts, he blogs and makes time to run several times a week — and, he says, he is having fun doing it all.

"I have to tell you one simple rule in any job. If you wake up for several days in a row and say, 'Why am I doing this?' then you're better off quitting," he says. "Not only has that not happened, I'm actually quite excited when I wake up every day. I just look at the day and hit it hard." ■

Apoorva Mandavilli *is a freelance science journalist based in New York and editor-in-chief of the autism news site SFARI.org.*

COMMENT

RAJIT SENGUPTA/CSE

Sunita Narain, director-general of the Centre for Science and Environment in New Delhi, calls for economical waste management.

Priorities for science in India

Ten Indian research leaders give their prescriptions, from better funding, facilities, mentoring and education to greater respect, fairness, autonomy and confidence.

PRADEEP P. MUJUMDAR
Share data on water resources

Professor of civil engineering, Indian Institute of Science, Bangalore

India is facing an imminent water crisis. Almost 100 million people have no access to safe drinking water, and most others experience regular shortages. More than one-third of the roughly 400 rivers that are monitored by the government are polluted.

Groundwater levels are alarmingly low in many areas, owing to overexploitation for irrigation and domestic supply. An estimated 60% of groundwater sources will be degraded in two decades. Cities consume vast amounts of energy to pump water over long distances from rivers and reservoirs, and unplanned urban growth is blocking drainage channels, causing flooding. Climate change will make matters worse. Water availability, demand and quality, as well as floods, droughts and salinity intrusion, will be affected.

But across India, hydrological research is hindered by a lack of access to good-quality data. The government bodies that are custodians of hydrological, meteorological, environmental and agricultural data are reluctant to share information openly. Combined with bureaucratic hurdles, this means that Indian researchers must either use poor-quality data or turn to US or European records.

To strengthen hydrological research and promote scientific decisions on water policy, the government must upgrade its data-collection, monitoring, communication and storage networks, in terms of both technology and density. The government's Water Resources Information System is an excellent start. Now it needs to provide real-time data on stream flow, soil moisture, groundwater levels and evapotranspiration.

'Critical zone observatories' that measure atmospheric, hydrological, biogeochemical, ecological and other fluxes in Earth's near-surface zone should be set up in each of India's seven hydro-climatic regions and ▶

integrated with others globally. Observatories should span different types of climate, terrain, demography, land use and land cover.

India needs multidisciplinary centres of excellence to address big questions — on water-system response rates to climate change, coupled forecasting of intense precipitation and floods, medium-range weather forecasts for agricultural water management and water contamination. These centres would also train the next generation of researchers to use holistic approaches. The Indian Institute of Science, Bangalore, has established such a centre this year. This step should be emulated nationwide with funding from government and private industry.

HIRIYAKKANAVAR ILA
Support the bulk of students

Professor of chemistry, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore

India's university system is broken. The higher-education system was started by the British in 1857 with the establishment of three universities — Calcutta, Bombay and Madras — and 28 affiliated colleges. After independence in 1947 and the creation of the University Grants Commission (UGC) in 1956, the number of universities grew exponentially. Today, there are more than 600, including about 200 private ones. Of the public universities, 46 are funded by central government the rest by state governments. A few, such as the central University of Hyderabad, do world-class research.

In addition to these, to improve training in basic sciences and technology, the government established 16 Indian Institutes of Technology and 5 Indian Institutes of Science Education and Research. There are also about 40 Council of Scientific and Industrial Research laboratories engaged in applied research, along with a few premiere research institutes, including my own.

The handful of these that compete at the international level contribute the bulk of high-quality scientific research in India. But they educate a tiny fraction of our students.

Facilities and teaching at the universities that serve more than 29 million students are alarming. Most are 'chalk and talk' classrooms with poor-quality teaching laboratories, let alone research laboratories. Faculty appointments are often made on the basis of political connections, caste or bribes, and funds are misappropriated. Inbreeding results: many highly qualified young scientists refuse to take up faculty positions in these universities because of the lack of infrastructure, the hostile environment and bureaucracy.

This is a disturbing situation. India needs trained, innovative minds to meet its formidable challenges. The state and central governments should take urgent action.

The government should appoint highly qualified, broad-minded vice-chancellors, who will recruit qualified faculty members and give them state-of-the-art research facilities with no external interference. Faculty should then focus on basic research and quality teaching, and encourage regional and international collaboration networks to strengthen scientific research. The government should also create many specialized research centres in the universities (like the CNRS in France). Fixing our university system will require a complete overhaul of the UGC, changes in institutional policy and legislation. This will be difficult with the present disconnect between science and policy in a government that has cut research budgets, focused on manufacturing and dissolved its scientific advisory committee.

YAMUNA KRISHNAN
Crack the cliques, enable visionaries

Professor of chemistry, University of Chicago, Illinois

To catapult India into the top five scientific nations, the country needs enabling policies that money can't buy. India has huge positives but it is hamstrung by socio-cultural issues, two of which I address here: a herd mentality and a paucity of early-stage mentorship. My ideas stem from my 15 years as a graduate student and young research-group leader in India.

Having recently moved from the National Centre for Biological Sciences in Bangalore to Chicago, Illinois, I have noticed a fundamental difference in the attitude of young US scientists from that of their Indian counterparts: their appetite for big problems.

"The country needs enabling policies that money can't buy."

'Going for great' is a skill acquired very early on in the West. Senior researchers spot gifted graduate students, connect them with the best scientific mentors, nurture them and ensure their visibility over decades.

In India, researchers generally start being mentored only when they show promise as young principal investigators. Thus a fresh returnee from a leading postdoctoral lab abroad suddenly becomes essentially invisible to key collaborators or contacts at home and elsewhere. This results in the returnee pursuing quality problems fragmented into smaller stories for more publications, but of lower visibility. The strategy is to edge slowly towards the big ideas. Often, these big ideas are either suddenly solved by counterparts in the West, or become outdated. A top-down, merit-based, long-term mentorship scheme — starting at the graduate-student stage — could prove transformative.

Cultivating excellence is a selective process that can be perceived as elitist. But India is trying its best to become an egalitarian society. It has some outstanding senior scientists — visionaries who care deeply about taking their nation from good to great. But their efforts are neutralized by a pedestrian majority intent on preserving the status quo.

Instead, these visionaries need to be empowered to take the tough decisions to make Indian science a meritocracy. We must take a census of researchers in all disciplines. Then, preserve scientists with research programmes of international standing, regardless of age, solely on the basis of performance during the past five years. Give them abundant support to ratchet up their programmes. Identify experienced scientists who could each nurture and mentor 5–10 emerging scientists and bring them up to an outstanding level. From such a platform, break open the moribund coteries that hold the system to ransom without themselves doing cutting-edge research. If this can be done, India will soon emerge as a scientific superpower.

I still bubble with optimism. India allows young people with the right attitude to thrive. The nation's history has many examples of the conscience of the majority successfully rejecting deeply embedded socio-cultural mindsets.

JOYASHREE ROY
Train more energy economists

Professor of economics, Jadavpur University, Kolkata

The energy sector will drive India's economic growth for the next three decades. Better access to electricity and cleaner fuel sources will enhance the population's health and well-being and boost industry. But the country faces major challenges, from implementing technologies on the ground to staying within global carbon-emissions limits while ensuring energy access for all.

The discussion so far is one-sided. In India, energy is seen mainly as a science-and-technology issue. There is money for developing microgrids and distributed power devices. But no serious research is being funded to examine the

A man cleans oil barrels for recycling in Dharavi, one of Mumbai's largest and oldest slums.

JONAS BENDIKSEN/MAGNUM

socio-economic impacts and influences.

How will distributed generation affect energy prices and social dynamics? What will happen when new actors such as suppliers of low-carbon energy and 'producer–consumers' enter the fray? Is there an optimum path — environmentally, socially and economically — for depleting natural resources?

India needs more energy economists. Energy has long been seen as an unfashionable topic in the country's universities, and few researchers specialize in the field compared with agriculture, trade, finance and the environment. India must create a forum of energy economists who can discuss and compare the models used to develop energy strategies and influence policy dialogues while understanding local nuances.

India is diverse, and political contexts matter. The energy sector, which meets basic service needs, is susceptible to partisan politics. But scientists have become distanced from policy-makers. Economists need to fill the gap by analysing which governance structures and regional cooperations might emerge under different energy-distribution scenarios and technological options.

A strategy to train the next generation of Indian energy economists could follow the model of the capacity-building programme for environmental economists, which ran from 1998 to 2003 through many participating universities and institutes, funded by the World Bank in collaboration with the then Ministry of Environment and Forests. Similar efforts are now being made by SANDEE, the South Asian Network for Development and Environmental Economics. Academics from around the world helped to train faculties in environmental economics, library content was improved, and grants and fellowships were offered so that Indian researchers could train overseas and build case studies in India.

Today, almost all universities in India have a well-defined, internationally comparable syllabus in environmental economics that is taught by well-trained teachers to plenty of students. It is now mandatory that an environmental economist be a member of each state's environmental impact assessment board. A similar approach for energy economics, starting with interested institutes, would encourage more researchers to seek solutions to India's energy problems.

RAGHAVENDRA GADAGKAR

Solve local problems

Professor of ecology, Indian Institute of Science, Bangalore; and president, Indian National Science Academy

Indian science suffers, today more than ever, from government apathy. This is exacerbated by the fact that India tries to run on the same track as the most developed countries and the best endowed institutes in the world. Only a handful of scientists and institutions in India can afford it, and then only by sequestering an unfair share of the country's scant funds. Even these players barely compete with their chosen peers — never really at the top, but in the 'also ran' category at best. This leaves most researchers and institutions with inadequate resources, and worse, feeling backward.

This is not the only model for success. If you cannot compete on the same track, you should try a different one. India should celebrate and encourage scientists who create their own research questions long before others make the topics fashionable, or those who bring different perspectives to existing problems. Most importantly, we should garland those who work on problems that are crucial to local contexts — even if they are of little interest to elite overseas universities or to 'high-impact' journals. Examples include endemic communicable diseases, groundwater contamination and traditional methods of biodiversity conservation.

India's systems for peer review, grants, publications, jobs, awards and fellowships punish any potential future leaders in such 'unsexy' fields. Instead, the country should develop new scientific ethics and etiquette. The research community should value, for instance, collaboration with small neighbouring colleges or universities instead of recognizing only international alliances. India should create a new peer-review system, a new ranking of journals and new measures of impact — all tailor-made for our needs, problems, diseases, natural resources and educational system. We need to believe in ourselves and not just chase world rankings — as individuals, as institutions and as a country. The enemy is within. So is the solution.

VINOD SINGH

Improve tertiary education

Director, Indian Institute of Science Education and Research, Bhopal

India produces around 9,000 PhD graduates a year in science and technology. This number sounds large, but for a population of about 1.3 billion it is not: the United States produces four times as many from a population one-quarter of the size. Moreover, the variation in quality of Indian PhD graduates and faculty members is a prime concern.

For India to be at the forefront of science and technology we need better governance systems for universities, institutes and research labs. We need more capable academics to provide leadership, nurture young talent and establish a superior research enterprise.

Indian universities are mired in bureaucracy. Archaic ordinances and rules set by the University Grants Commission have stifled the spirit of academic excellence and hampered institutions' flexibility. A lack of passionate leadership coupled with poor funding has blunted their edge.

Leading the way are premier government-funded centres such as the 16 Indian Institutes of Technology, the Indian Institute of Science in Bangalore, the Tata Institute for Fundamental Research in Mumbai, and the 5 Indian Institutes for Science Education and Research. These have one academic

director, who reports to a board of governors of eminent academics, researchers and industrialists. An effective leader — with excellent research and administration skills — can cut through bureaucracy. Other public universities should similarly be made autonomous.

Centrally funded laboratories, tasked with industrially relevant research, should be run along similar lines and integrated with nearby universities and institutes. This would strengthen applied and interdisciplinary research.

In 2009, the Science Engineering Research Board was created to make government science funding quicker and fairer. Its performance now needs to be benchmarked against overseas granting agencies such as the US National Science Foundation.

Quality-control mechanisms must be established for the national accreditation and assessment of Indian PhDs and to improve research and educational training. Doctoral fellowships and research funds should be created in areas of national priority, including food security, energy and the environment. It is high time that India fixed its tertiary education system.

UMESH VARSHNEY

Make science an attractive career

Professor and chairman, Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore

Is there a dearth of talent in India? Certainly not. Is there a dearth of unstoppable achievers and innovators? Yes: because making talent shine takes a culture that is proud of its scientists and a charged intellectual environment that nurtures, mentors and drives them. The efforts made by a handful of educational institutions, academies and a few others are crucial — but inadequate. We must halt the deterioration in higher-education standards in hundreds of universities, which train and produce huge numbers of science undergraduates and graduates.

Science graduates are deprived of meaningful practical training because of poor funding, government interference, inappropriately recruited faculty members and a lack of laboratory facilities in most of these centres of learning. At this crucial stage in their careers, students are missing out on the mentoring required to instil the culture of science and the habit of analytical thinking and questioning. And once scientists are trained? They work with inadequate, ill-maintained equipment, and in isolation from stimulating peers, being so few in number and so geographically dispersed.

It is imperative that the universities that produce the largest numbers of science graduates are revived so as to be capable of contemporary research. The process can be difficult and slow, or expensive and experimental. One such experiment would be to fund science generously. Another related one would be to pay researchers enough to make science a socially acceptable profession.

Meanwhile, the resilient among us must continue by *jugaad* — the characteristically Indian technique of making do — to try to maintain the scientific base that exists. If only the management of science were left to scientists, India could put its research on the world map — just as it put the Mangalyaan probe into orbit around Mars.

The daily delivery of drinking water causes frenzy in Delhi.

JONAS BENDIKSEN/MAGNUM PHOTO

KRISHNA N. GANESH

Connect research with education

Director, Indian Institute of Science Education and Research, Pune

Historically in the Indian education system, faculty members who teach undergraduates do not do research, and those involved in research (in national laboratories and universities) do not teach undergrads. This is the opposite of the conventional Western university system.

To inject research-led undergraduate teaching, five Indian Institutes of Science Education and Research (IISERs) were set up between 2006 and 2008, in Pune, Kolkata, Mohali, Bhopal and Thiruvananthapuram; the sixth one is being established this year in Tirupati. At the IISERs, students are exposed to research early in their careers, in state-of-the-art labs. Customized curricula connect theory taught in the classroom with lab experiments. Courses in social sciences, ethics and science communication round out the education.

This alliance of education and research catapulted the IISERs to fourth place in India in the 2014 Nature Index, which ranks institutes' outputs — no mean feat for institutes less than a decade old. Together, the IISERs now have 350 faculty members and 3,500 students and will reach their final capacities (2,000 students and 200 faculty members per institute) by 2019.

However, Indian research institutes still fare poorly in global rankings in terms of publication quality. They must try to attract international visiting faculty members and research students, and establish good ties with industry. More than 60% of the 600 students who have already graduated from the IISER system have secured PhD positions in leading universities abroad.

This sort of brain drain is why the Indian system is seriously afflicted by a lack of postdoctoral fellows, who are the engine of the research enterprise elsewhere. Even the best professors in India depend mainly on PhD students for their research. The government's proposed fellowship plan to send Indian PhD holders abroad to gain experience and training in emerging areas should be converted to a programme that 'twins' Indian institutions with foreign research centres, with candidates spending half their time in India. Fellowships could also be opened to foreign nationals wishing to work in India. To assure career progression, these should dovetail into existing tenure-track systems — such as the INSPIRE Faculty Scheme, the University Grants Commission Faculty

Recharge Programme and the Ramanujan and Ramalingaswami fellowships.

To retain or attract back our best young scientists, and entice industry investments, India must create advanced research facilities and assured and scalable research funding, and must foster supportive mentors and visionary institutional leaders. To realize all this, the highest-achieving institutions must be granted immunity from general budget cuts and endowed with 20–30% more in funding for the next ten years, in autonomously controlled budgets. Germany's Max Planck Institutes provide an ideal governing model.

This year has seen cuts in the proportion of gross domestic product spent on science and technology, from an already low starting point of 0.9%. This risks not only undoing the progress achieved, but also doing irreversible damage. At the same time, many important scientific agencies including the Department of Science and Technology (until recently), the Council of Scientific and Industrial Research, the Indian Council of Medical Research and several national laboratories have been without chiefs for more than a year, which has stalled strategic decision-making and dented morale.

In the absence of the Science Advisory Council to the Prime Minister, there is no channel for enlightening the government on the crucial role that scientists could have in addressing India's growing challenges in energy, health, environment, water and education. The country's science academies must build such a bridge. India has a vast supply of talented young people; it is our duty to nurture and harness their talent for a better tomorrow.

SUNITA NARAIN

Manage waste frugally

Director-general, Centre for Science and Environment, New Delhi

India has a huge waste problem. Untreated sewage is defiling rivers and water bodies; industrial chemicals such as cadmium and nitrates are seeping into the ground and polluting the air; and solid waste from kitchen scraps and plastic packaging is piling up in our cities. The problem requires more than management. We need innovative and realistic solutions that match our pockets and our regulatory and governance abilities.

Take sewage. Flushing excreta down toilets is expensive and resource intensive. It uses water as both the carrier and the final dumping point. This approach works in countries that have the means to build huge water-supply and retrieval infrastructures and to pay for maintenance and upgrades of technologies to manage and treat pollutants — from biological waste to toxins. It does not work in India, where there are limited funds for supplying essential services to more than one billion people. A country that is poor — but fast becoming richer and more wasteful — needs a whole new paradigm.

The key obstacle is that everyday challenges are not top priorities for research and innovation. Indian science has always been fascinated by the 'masculine' agendas of space and genetics, not reinventing the toilet.

Instead, science must meet the needs of poor people. We need to devise ways to prevent pollution rather than cleaning it up afterwards. Indian research has to be more humble, nimble and investigative. It has to learn from its poorest and most illiterate people: how they cope with scarce and diverse resources by being frugal and in tune with their environment.

India's ambition should be to become the front-runner in the race to save the planet.

Naba Mondal, director of the India-based Neutrino Observatory project.

NABA K. MONDAL

Build big physics facilities

Senior professor, Tata Institute of Fundamental Research, Mumbai

India has an illustrious history in high-energy physics. But two factors make me worry that it will struggle to maintain its position: a shortage of instrument builders and the difficulty of getting planning permission for big facilities.

Technological advances lie behind breakthroughs in particle physics. Indian scientists' enthusiasm and skill for building particle detectors put them at the forefront of the field early on. In the 1950s and 1960s, Indian physicists pioneered experiments with cosmic rays, and developed cloud chambers for use at high altitude. The first published detection of atmospheric neutrinos was made in 1965 with an instrument installed in a mine at Kolar Gold Fields (KGF) in Karnataka state. The first dedicated experiment to study proton decay was carried out at KGF in the early 1980s.

Today, there is little enthusiasm among India's young researchers for building instruments. One reason is that the pay-off is years in the making: researchers lose out in terms of publications compared to peers working in the lab or doing theoretical research. They find it difficult to compete in the academic job market.

Unless we devise metrics that recognize instrument development and retain these skills, it will be difficult to host high-energy-physics experiments in India. India's participation in international projects will be limited to data analysis, making us unequal partners.

Another obstacle is the slow and complex approval procedures for large experimental programmes in India. This is compounded by widespread opposition to large-scale projects by political opportunists and activists on flimsy grounds. In a healthy democracy, meaningful debates are welcome. In India, they are increasingly becoming indiscriminate and adversarial.

For example, controversy has broken out over the proposed India-based Neutrino Observatory, an underground lab in Tamil Nadu for research on neutrinos and related particle physics. The project, of which I am director, received government approval in December 2014. To stay globally competitive, it needs to be up and running by 2020. But we are far from breaking ground. By spreading fictitious fears about neutrinos, a small local political party and a handful of activists have sowed doubts in the minds of local people and made it extremely hard for us to get the required planning permissions.

Unless scientists speak up collectively, it will be prohibitively difficult to develop major science infrastructure in India. ■

SUZANNE LEE/PANOS

Kamla Devi, Rajasthan's first female solar engineer.

Rethink India's energy strategy

Address the needs of poor and rural households, target subsidies and support low-carbon industries, urge **Arunabha Ghosh** and **Karthik Ganesan**.

India's policy-makers have three big energy goals: providing everyone with access to energy, securing energy supply and trying to limit carbon emissions without encumbering the nation's growth. These important concerns miss the point.

Energy access cannot be assured by progress towards a simple target such as supplying power 24 hours a day, 7 days a week, nationwide. India has deep divides in the quantity and quality of energy consumed across income groups and between rural and urban households. Fuel subsidies are poorly designed and the strategies to reduce them to enhance energy security are heavy-handed. And because of limited action by the world's largest emitters, there is little left in the global carbon budget before planetary safety limits are breached. Clean energy and alternative growth is imperative.

India's energy priorities should be reframed as follows: to cater to the different energy demands of citizens of various economic strata; to direct energy subsidies to benefit the poor; and to promote low-carbon industry.

DISPARATE DEMAND

Urban India aspires to have a reliable 24/7 electricity supply — voltages currently drop at peak demand times such as during evenings. Meanwhile, more than one-third of India's households, mostly poor and rural, are not connected to the electricity grid. For those that are, blackouts last 4–16 hours a day. The poorest households consume one-quarter of the energy of those at the highest income levels. Urban centres are in effect subsidised by rural areas, which are being overcharged for substandard service[1]. The poorest households pay 30% more per unit of useful energy than the richest[2].

One solution to these disparate demands is to deliver more electricity through the grid while adopting cleaner energy sources. The Indian government has announced ambitious plans for renewable energy: up to 175 gigawatts (GW) of installed capacity by 2022. There are many challenges to achieving this target, from the availability of resource data on which to base decisions and managing risks to the high cost and the huge variability across the grid in terms of energy sources and infrastructure.

Meanwhile, the promise of reliable electricity through centralized infrastructure

and systems remains unfulfilled. This is in part because most electricity utilities suffer financial difficulties — they lost more than US$19 billion in 2012–13 (ref. 3). One solution is to tap smaller-scale distributed renewable energy sources, primarily solar, biomass and small-scale hydropower. Off-grid power based on these technologies has advantages such as network resilience, flexibility and environmental and health benefits[4].

More than one million households in India rely on solar off-grid systems for lighting. A further 20 GW of energy capacity could be achieved if 15% of irrigation pumps were converted to solar energy. Renewable-energy applications can provide heating, cooling, cooking and mechanical power as well as electricity[5].

More than 250 companies across India, with long supply chains and networks of village-level entrepreneurs, operate in the decentralized clean-energy sector already. They demonstrate that putting power in the hands of poor people can begin a transformation in how energy access is understood and delivered. At the same time, such rapid growth and geographical spread could result in variable quality of service and expensive energy for poor people. More training would help to keep up standards.

The challenge is to balance two types of investment: those in the centralized grid, which are key to the aspirations of millions of higher-income households, and funds for standalone systems in isolated and underserved areas or for integrating such systems to the grid.

RATIONAL SUBSIDIES

Another reason for pursuing renewable energy in India is to avoid the pitfalls of a growth strategy mostly based on fossil fuels. Already, imports account for more than 80% of India's crude oil and 25% of its coal and gas, raising worries about supply and price volatility[6]. Petroleum constitutes nearly 30% of all commodity imports, leaving India little fiscal room to shrink its large current account deficit.

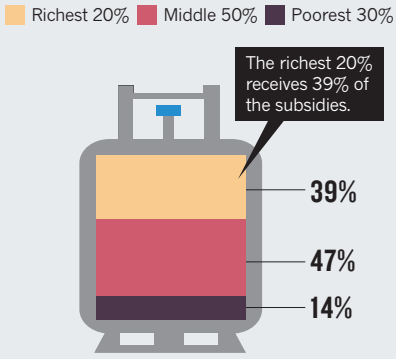
India hands out generous energy subsidies, most of which are not means-tested (see 'Energy imbalance'). For example, in 2013–14 the government gave away $8 billion worth of subsidies for liquified petroleum gas (LPG)[7]. Yet less than half of urban households and only 6% of rural ones exclusively use LPG for cooking. Traditional biomass fuels such as wood account for 20% of Indian households' energy use. The government must rationalize subsidies and target them better.

"The poorest households pay 30% more than the richest per unit of useful energy."

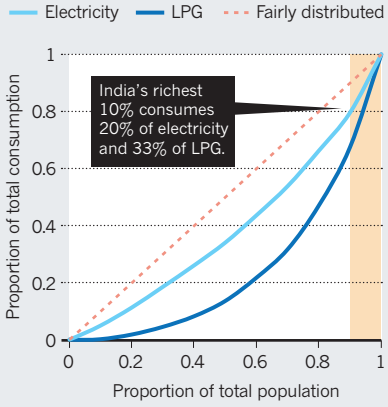
ENERGY IMBALANCE

Liquified petroleum gas (LPG) fuel is heavily subsidized by the government, even though it is used mainly by high-income families. A rural electrification programme started in 2005 has improved the fairness of electricity consumption.

1 **Share of LPG subsidy, by income bracket**



2 **Energy-consumption inequality**



SOURCE: REF. 7

A well-designed programme would increase access to modern cooking energy (electricity and gas) for the same budget. For instance, reducing subsidized LPG to 9 cylinders (instead of 12) per year per connection could save the government $724 million. Excluding the richest 15% of households from the subsidy could save $1.18 billion annually. The savings should be redirected to increasing the availability in rural areas of cleaner cookstoves and biogas, and could extend LPG provision to 30 million more households.

WHAT TO MAKE IN INDIA?

Energy and climate policies are closely tied to industrial policy. Even on a low-carbon energy pathway, total primary energy consumption in India will at least double by 2030 (compared to 2011 levels). Energy efficiency alone — in industry, residential and commercial spheres — cannot mitigate climate change.

Although unemployment rates in India are low (less than 5%) nearly 35% of employment is casual labour. The government's Make in India campaign, launched in September 2014, calls for aggressive job creation through rapid growth in the industrial sector.

Manufacturing consumes nearly one-third of India's primary energy supply, and contributes to 16% of gross domestic product (GDP) and more than 20% of direct emissions[8]. These emissions would grow, should India achieve its target of 25% contribution to GDP from manufacturing.

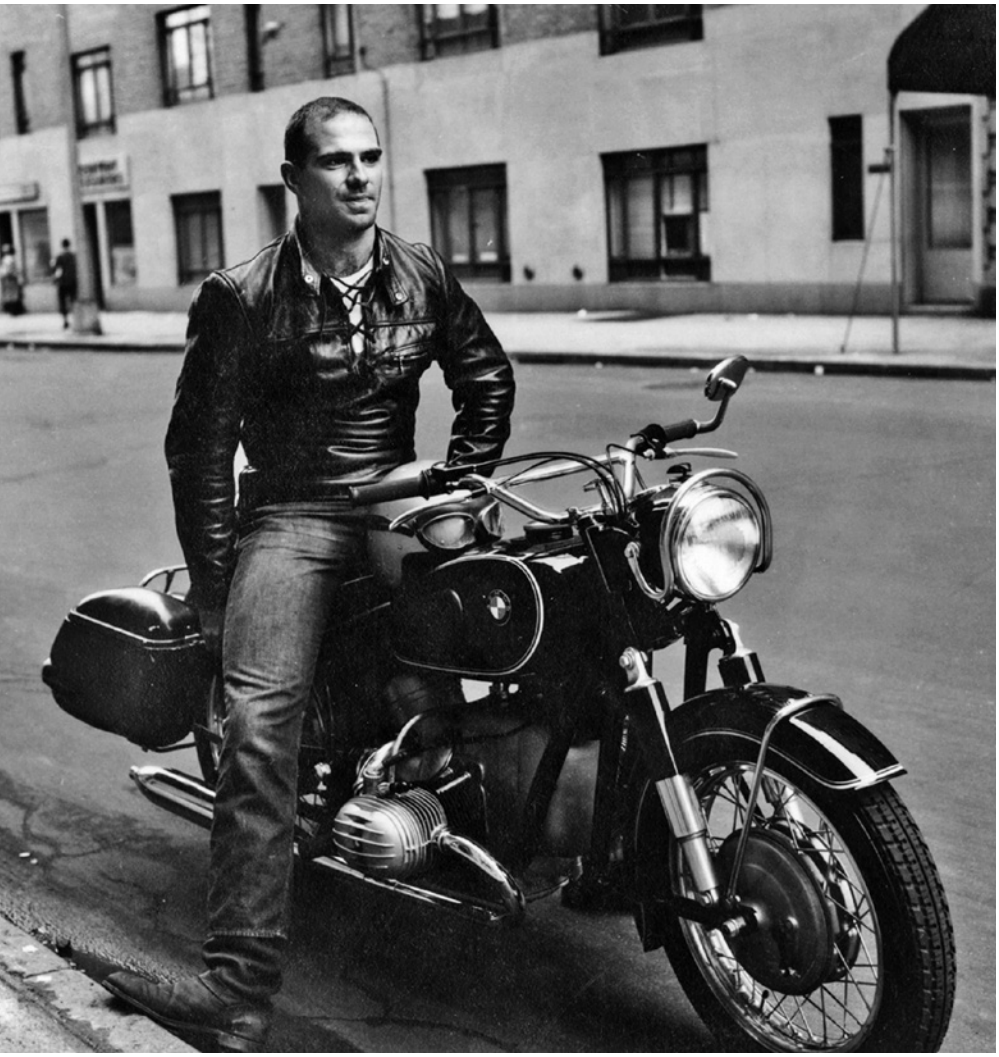
The best opportunity for decarbonization, therefore, is the power sector — which contributes nearly 38% of overall emissions[8]. Here, renewable energy could account for about 30% of the electricity mix by 2030.

In sectors such as metal production, non-metallic minerals, chemicals and textiles, which contribute most to manufacturing GDP, fuel accounts for 9–23% of all input costs compared to the industrial-sector average of 5%. Energy efficiency and alternative fuels should play a key part in decarbonizing these sectors. India's cement industry, for instance, is one of the world's most efficient, but it also accounts for 7% of the country's emissions. Here, technological advances such as refuse-derived fuels could save 600 million tonnes of coal, 550 billion units of electricity and 3.4 gigatonnes of carbon dioxide emissions between now and 2050.

A shift to a different industrial mix is required: away from such energy-intensive sectors and towards metal fabrication, manufacture of computers and electronics, electrical and mechanical machinery, advanced materials, biotechnology, nanotechnology and photonics. This would lower the energy footprint of India's industrial growth. ■

Arunabha Ghosh *is chief executive and* **Karthik Ganesan** *is senior research associate at the Council on Energy, Environment and Water, New Delhi, India.*
e-mail: arunabha.ghosh@ceew.in

1. Harish, S. M. & Tongia, R. *Do Rural Residential Electricity Consumers Cross-Subsidise their Urban Counterparts? Exploring the Inequity in Supply in the Indian Power Sector* (Brookings, 2014).
2. Ganesan, K., & Vishnu, R. *Energy Access in India — Today, and Tomorrow* (Council on Energy, Environment and Water, 2014).
3. Power Finance Corporation. *The Performance of State Power Utilities for the years 2010–11 to 2012–13* (PFC, 2013); available at http://go.nature.com/io3psm
4. Kammen, D. M., Alstone, P. & Gershenson, D. *AIP Conf. Proc.* **1652,** 14 (2015).
5. WWF-India and CEEW *RE+: Renewables Beyond Electricity* (WWF-India and CEEW, 2013).
6. Steven, D. & Ghosh, A. in *The New Politics of Strategic Resources: Energy and Food Security Challenges in the 21st Century* (eds Steven, D., O'Brien, E. & James, B.) 40–70 (Brookings, 2014).
7. Jain, A., Agrawal, S. & Ganesan, K. *Rationalising Subsidies, Reaching the Underserved: Improving Effectiveness of Domestic LPG Subsidy and Distribution in India* (Council on Energy, Environment and Water, 2014).
8. Indian Network for Climate Change Assessment. *India: Greenhouse Gas Emissions 2007* (Ministry of Environment and Forests, 2010).



DOUGLAS WHITE

Oliver Sacks in New York in 1961.

AUTOBIOGRAPHY

In search of self and science

Tim Radford revels in Oliver Sacks's memoir of his youth as a biker, druggie, muscle-builder — and scientist.

A young man sets out to find himself. He discovers motorbikes, leather, speed and thrills, lives on a kibbutz and has a learning experience with a prostitute in Paris. (The climax is a shared pot of lapsang souchong tea.) In the United States, as a young medic, he takes to the road on his motorcycle: *Easy Rider* in the landscape of John Steinbeck. He also takes to marijuana, LSD, methamphetamine and morning-glory seeds. He has the appetite for science but not the patience for research. He is drawn to the helpless, the hopeless and the lost, but manages to annoy the hell out of his peers, so he drifts from job to job.

This is the stuff of a certain kind of mid-twentieth-century novel. It is also the youth chronicled by Oliver Sacks in what might be his final reminiscence, *On the Move*. If it is the last, it is the coda to an astonishing life: Sacks is a scientist, a doctor of medicine and a clinical consultant who has also had a brilliant career as a best-selling author and man-about-neuroscience.

On the Move: A Life
OLIVER SACKS
Alfred A. Knopf: 2015.

None of Sacks's journey into mythic America was planned. The enthusiastic biker who first crossed the Atlantic from Britain in 1960 with vague dreams of joining the Royal Canadian Air Force or becoming a lumberjack instead achieved enduring recognition and status as professor of neurology at New York University. But you might not predict it from this account of uncertain beginnings and peripatetic adventures.

Sacks, who had been a junior doctor at the Middlesex Hospital in London, became an intern of uncertain migrant status working with neurosurgeons at Mount Zion Hospital in San Francisco. He fell in love with California, and hung around Muscle Beach in Santa Monica, birthplace of the US gym revolution. Here as elsewhere, he did nothing by halves, consuming "five double cheeseburgers and half a dozen milkshakes per evening" to bulk up for his power lifts. He broke a weightlifting record ("a squat with a 600-pound bar on my shoulders") and several bones. His capacity for alcohol matched his appetite for learning (at around 17, deep in James Joyce's titanic 1922 novel *Ulysses*, he sipped his way through a litre of aquavit during the North Sea ferry crossing from Norway) and his ability to connect with others.

That talent was to inform his understanding of patients, colleagues and readers. His friends have ranged from truck drivers and Hells Angels to polymath-director-scientist Jonathan Miller, geneticist Francis Crick, poet W. H. Auden — and Hollywood star and muscleman connoisseur Mae West, who chatted him up while he was moonlighting at a Los Angeles hospital.

At the heart of this picaresque adventure is an unhappy secret. When Sacks admitted in 1951, aged 18, that he might prefer boys to girls, his mother called him "an abomination" and wished that he had never been born. However, nothing 'abominable' had yet happened. That was to hit several years later, when, determined to lose his virginity, he headed to Amsterdam and its gay bars, but drank so much gin that he was unconscious during his deflowering.

Sacks was not without connections and luck — a loving medical family in London, a scholarship to the University of Oxford, cousins including iconic US cartoonist Al Capp and Abba Eban, the Israeli scholar and diplomat. After the success of Sacks's first book, *Migraine* (Vintage, 1970), his father joked that he no longer spoke of himself as "Abba Eban's uncle" but as "the father of Oliver

➲ NATURE.COM
For more on Sacks in the *Nature Podcast*:
go.nature.com/awsgzg

Sacks". And can he write? Over his life, Sacks has filled 1,000 notebooks and journals, not counting journalism, medical notes and a lost suitcase full of photographs and notes — written in bars and restaurants, up mountains and in airports. He has more than a dozen books in print. Harold Pinter wrote a play inspired by his second book, *Awakenings* (HarperPerennial, 1973); Penny Marshall directed the film. *Awakenings* also inspired a ballet, and Peter Brook directed a French theatre production of *The Man Who Mistook His Wife For A Hat* (Summit, 1985). Michael Nyman wrote an opera on the same work.

Some of *On The Move* feels ripe for a US heavyweight such as the novelist James Baldwin. Other parts are untidily told, padded with extracts from letters home or the young adventurer's first attempts at writing. It is not quite clear how that youthful, ready-for-anything medic metamorphosed into a distinguished professor. We piece the story together from anecdotes of foolhardy adventure and episodes of clinical encounter. Sacks writes about people with migraines, Tourette's syndrome or Parkinson's disease, autism, epilepsy, colour blindness, serious mental illness and the post-encephalitics of the Beth Abraham Hospital in the Bronx, New York, who are the subjects of *Awakenings*. These are the stuff of his books: not just medical cases, but warm, quirky and aware.

"This is a compelling front-line dispatch from half a century's wonderful exploration of brain, mind and nervous system."

This is another compelling front-line dispatch from half a century's wonderful exploration of brain, mind and nervous system. It is a valedictory memoir, and one with a tentatively happy ending. At 76, this lonely writer ("it has sometimes seemed to me that I have lived at a certain distance from life") found enduring love. But the book's text was handed to the publisher before Sacks, now 81, was diagnosed with cancer of the liver. He has just written about that in *The New York Review of Books*, and of — in the words of Friedrich Nietzsche — "a reawakened faith in a tomorrow and the day after tomorrow". Here's hoping there may yet be an epilogue. ■

Tim Radford *is a former science editor of* The Guardian*, and author of* The Address Book: Our Place in the Scheme of Things.
e-mail: radford.tim@gmail.com

MARK GARLICK/SPL

SCIENCE FICTION

After the cataclysm

John Gilbey delights in a vast, technologically charged tale from a science-fiction supremo at the top of his game.

Earth undergoes a catastrophic transformation in *Seveneves*.

Neal Stephenson writes big science fiction, literally and figuratively. Weighing in at some 900 pages and stretching nearly 5,000 years into humanity's future, *Seveneves* is no exception.

It traces an epoch in which humankind and the environment change profoundly. The bulk of the novel is the lead-up to, and immediate aftermath of, a stunning cosmic event that leaves humanity teetering on the edge. The remainder describes a renaissance with only faint echoes of what we recognize as human culture.

The cataclysm is the destruction of the Moon by a mysterious agent. As Earth is assaulted by a rain of debris from the shattered satellite, the vast majority of the human population faces oblivion. The core of the story relies on current, or currently anticipated, technologies — weaving a plausible tale of how a tiny number of survivors, the "seveneves" of the title, might secure a future for our species. Stephenson imagines the rebirth as a division into seven races, based on the genetic profiles of the founders. The future cultures have both old and new social problems, but also fresh insights and resources with which to address them.

The epic injury to Earth looms in the very first sentence: a masterful attention-grabber. Stephenson maintains tension and energy, as well as a remarkable technical complexity, both literary and scientific. I repeatedly found myself sketching parts of the dramatically scaled mechanical constructs that enable later stages of the story — such as whip-like machinery to capture high-flying gliders and transfer them to Earth orbit — to judge whether they were feasible. They were.

Seveneves
NEAL STEPHENSON
William Morrow: 2015.

Comparisons with other sci-fi epics are inevitable. The Culture series by Iain M. Banks carries similar social and sexual complexities, massive terraforming machinery and off-world habitats, and shares Stephenson's delight in clever characterization and off-beat humour. Arthur C. Clarke's *The Fountains of Paradise* (1979) embodies related technical solutions, and has a postscript that makes a temporal and socio-biological leap of the same scale. Olaf Stapledon's classic *Last and First Men* (1930) paints a similarly portentous picture of genetic manipulation, cosmic cataclysm and the potential future forms of humanity — albeit with a massively larger scope and extending forwards for millions of years. But what distinguishes *Seveneves* for me is Stephenson's handling of the characters. There is an almost Malthusian detachment in how he introduces, builds, then violently dispatches characters who in novels with less robust reasoning would be saved by a clever plot device.

This is hard sci-fi in a real and welcome sense, ruled by unremitting physical laws, unlike the negotiable rules of the action thriller. People die because their deaths are inevitable, and many pass unremarked because the disaster's scale is so vast. Their sacrifice is tied to the theme of engineering the survival of the human race. Science fiction often suffers from a disparity between the impressive scale of the scenery, and the size of the characters and how they are developed. Stephenson balances these aspects well, avoiding cookie-cutter scientists and the all-too-common characterization of technologists as brilliant but conflicted renegades.

I did find myself mulling over the casting for the film that is sure to follow. Someone needs to talk to Morgan Freeman's agent, that's all I'm saying. And an almost throwaway early scene is never quite resolved, making it clear that there is significant scope for sequels. I very much hope that Stephenson is working on them. ■

John Gilbey *is a science and science-fiction writer. He teaches in the department of computer science at Aberystwyth University, UK.*
e-mail: gilbey@bcs.org.uk

NEUROPHYSIOLOGY

The man who bared the brain

Alison Abbott encounters the discoveries of Renaissance anatomist Andreas Vesalius.

It is only when you read the words that Andreas Vesalius wrote as an angry young man in the 1540s that you get a feeling for what drove him to document every scrap of human anatomy his eye could see. His anger was directed at Galen, the second-century physician whose anatomical teachings had been held as gospel for more than a millennium. Roman Empire law had barred Galen from dissecting humans, so he had extrapolated as best he could from animal dissections — often wrongly.

Human dissections were also banned in most of sixteenth-century Europe, so Vesalius travelled to wherever they were allowed. He saw Galen's errors and dared to report them, most explicitly in his seven-volume *De Humani Corporis Fabrica* (*On the Fabric of the Human Body*), which he began aged 24, working with some of the best art professionals of the time. His mission to learn through direct and systematic observation marked the start of a new way of doing science.

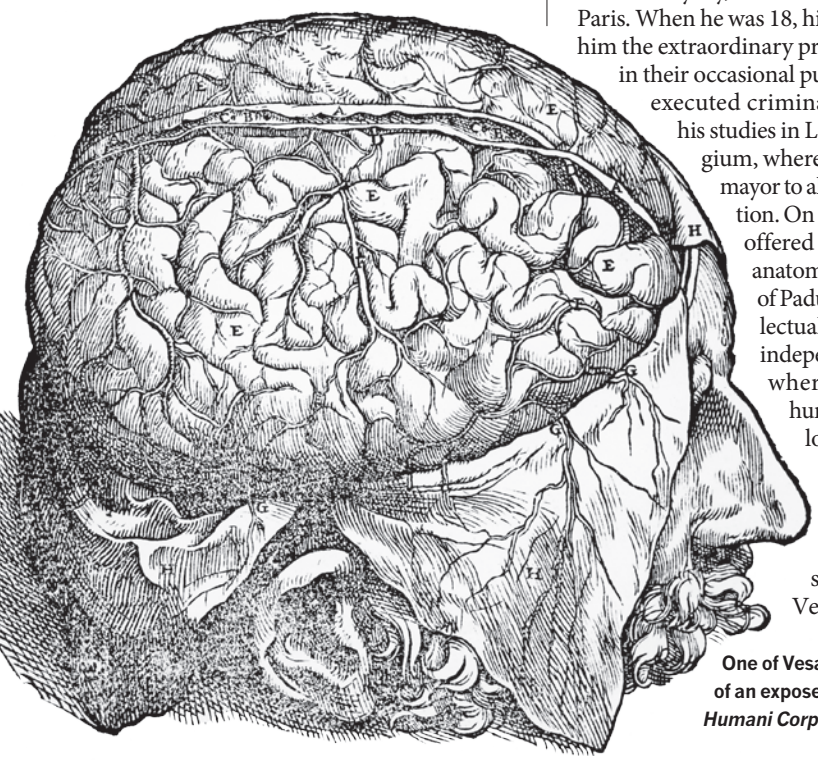
In *Brain Renaissance*, neuroscientists Marco Catani and Stefano Sandrone present a translation from the Latin of the *Fabrica*'s last volume, which focuses on the brain. Through it we can appreciate Vesalius's extraordinary attention to detail, and his willingness to believe his eyes, even when what he saw contradicted established knowledge. We learn his anatomical vocabulary. For example, he called the rounded surface protuberances near the brain stem "buttocks" and "testes"; these are now known as the inferior and superior colliculi, or 'little hills', which process sound and vision.

We hear Vesalius — who died at 49 in 1564, the year Galileo was born — berating Galen and his sixteenth-century followers with unrestrained sarcasm. But we also see a man not quite prepared to put into plain words the theological consequences of some of his discoveries, particularly his failure to find an anatomical explanation for the 'human spirit' in the brain.

Brain Renaissance is not the only English translation, but it is the only one available at a price that individuals might afford. Accompanying texts by Catani and Sandrone place the work in its historical and scientific context; a biography of Vesalius is rich in elements familiar to scientists today, such as the fear of plagiarism and pernicious academic rivalry. And a brief final section on the history of neuroscience warns against the temptation to move away from direct observation into overly abstract theory.

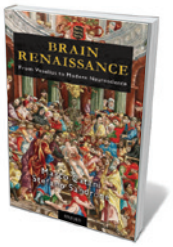
Born into a well-to-do Brussels family of physicians and pharmacists, many of whom attended royalty, Vesalius studied medicine in Paris. When he was 18, his teachers allowed him the extraordinary privilege of assisting in their occasional public dissections of executed criminals. He continued his studies in Leuven, now in Belgium, where he persuaded the mayor to allow human dissection. On graduation, he was offered a professorship in anatomy at the University of Padua, Italy — an intellectual hotbed politically independent of the Pope, where the practice of human dissection was long established.

Padua is close to Venice, which was home to important schools of artists. Vesalius recruited members of Titian's workshop to attend his dissections and provide the *Fabrica*'s exceptional illustrations. In the first two volumes, skeletons and flayed figures pose in romantic landscapes full of classical iconography. Figures in the other volumes are less ornate, but clear and fine. The brain is often shown encased in the skull with the top removed, revealing cross-sections at different depths; other images depict individual brain structures such as the cerebellum.



One of Vesalius's illustrations of an exposed brain from *De Humani Corporis Fabrica*, 1543.

BIBLIOTHÈQUE DE LA FACULTÉ DE MÉDECINE, PARIS/ARCHIVES CHARMET/BRIDGEMAN IMAGES



Brain Renaissance: From Vesalius to Modern Neuroscience
MARCO CATANI AND STEFANO SANDRONE
Oxford Univ. Press: 2015.

Vesalius's thinking was as influenced by prevailing technologies as ours is, and as Galen's was, note Catani and Sandrone. Whereas today we draw on computer and social networks for metaphors about brain function, for Vesalius and Galen the technology of reference was hydraulics, which almost miraculously kept the waterways and plumbing systems of their cities functioning. Both saw the brain in these terms, with the functional units being the liquid-containing ventricles rather than grey and white matter. Galen held that the vivifying force of the *pneuma physicon*, or 'animal spirit', flowed down through the ventricles, then through hollow nerves to nourish all parts of the body. Vesalius ruled this out on anatomical grounds: he showed that there is no physical outlet through the skull. But he still searched for flow routes that would, for example, funnel 'brain phlegm' into the nostrils.

Vesalius was aware of the value of his work, and of the academic jealousies that could work against him. He chose not to use a Venetian printer who was producing a rival, Galen-based anatomical tome, perhaps because he feared that the printer would deliberately delay publication of his own study. Instead he crossed the Alps to Basel, Switzerland, where, still fearing intellectual theft, he stayed to oversee the printing. He whiled away his time by boiling the body of an executed murderer to get at his bones. The reassembled skeleton is still displayed in the city's university. ■

Alison Abbott *is* Nature*'s senior European correspondent.*

Correspondence

IPCC: social scientists are ready

Social scientists are ready to work as full partners with physicists and ecologists on climate-change assessments by the Intergovernmental Panel on Climate Change (IPCC), government agencies and other organizations (D. Victor *Nature* **520,** 27–29; 2015).

Social scientists are already involved in climate discussions to some extent through societies of various disciplines (see, for example, go.nature.com/cixl9y). For more than 20 years, the US Board on Environmental Change and Society and the former International Human Dimensions Programme on Global Environmental Change have been examining the implications of social-science research for global environmental change — and vice versa (see, for example, go.nature.com/24t7mf).

However, social scientists can do more to engage with climate change by applying their disciplines and investigating human–environmental interactions that have their own dynamics. Further investment will help to develop integrated social and environmental data sets to support these analyses. These efforts will allow social scientists to test and expand the scope of their concepts and methods to contribute important insight into such relatively new issues.

Paul C. Stern *National Research Council, Washington DC, USA.*
Thomas Dietz *Michigan State University, East Lansing, Michigan, USA.*
pstern@nas.edu

IPCC: calling social scientists of all kinds

The invitation to contribute more to the Intergovernmental Panel on Climate Change (IPCC) should not be limited to social-science fields that are immediately relevant to decision-making processes, as David Victor seems to imply (see *Nature* **520,** 27–29; 2015).

A wider mix of social scientists can more effectively contribute valuable knowledge to the often-contentious societal and policy issues around climate change. They will help to retain and strengthen the role of the IPCC in international policy-making and public discourse by using qualitative and quantitative, theoretical and empirical, basic and applied approaches.

This should not be difficult: most international social-science associations have working groups on climate and environmental change, and integrated science platforms such as Future Earth involve hundreds of social scientists worldwide (www.futureearth.org). These social scientists should be equal partners in the IPCC's framing and scoping processes, along with their natural-scientist colleagues and policy-makers.

Mathieu Denis *International Social Science Council, Paris, France.*
Susanne C. Moser *Future Earth, Santa Cruz, California, USA.*
mathieu@worldsocialscience.org

Plant identification is key to conservation

Isabel Marques' call to update traditional botany teaching beyond plant morphology seems to devalue the importance of taxonomy and systematics (*Nature* **520,** 295; 2015). The identification of new plant species is still as relevant today as the discovery of new genes and gene functions — and is crucial for conservation efforts in developing countries.

Enormous numbers of plant species in Brazil and Madagascar, for example, still await formal description. The skills needed to meet this challenge and their ability to attract funding should not be dismissed. The Brazilian government is already providing funding, although it will be a long time before the Malagasy government can do so.

The botanical community is fast filling the gap between herbarium work and metagenomics, despite dwindling funding (see also M. Kemler *Nature* **521,** 32; 2015).

A good place to start reviving interest in a botanical education, and hence strengthen this community, would be to include striking plant species such as *Rafflesia* and *Sarracenia* in televised nature programmes, rather than focusing on charismatic megafauna.

Anna Trias-Blasi, Maria Vorontsova *Royal Botanic Gardens, Kew, UK.*
a.triasblasi@kew.org

Citizen science is not enough on its own

Citizen scientists' important contributions to biodiversity conservation are constrained by their focus on data collection and public outreach in wealthy, accessible places. Sustainable conservation actions require initiatives such as those supported by the Participatory Monitoring and Management Partnership (www.pmmpartnership.com), in which data collected by land owners and resource users help to guide local decision-makers on conservation management.

Citizen scientists do not formulate research questions, analyse data or implement management solutions on the basis of research findings. By contrast, participatory monitoring by local and indigenous communities in tropical, Arctic and developing regions enables them to propose solutions for environmental problems, advance sustainable economic opportunities, exert management rights and contribute to global environmental data sets.

Such monitoring could benefit from the large-scale databases and knowledge integration pioneered by citizen science. Conversely, citizen science could benefit from the community-based monitoring practices used to build data-collection methods, analytical tools, communication networks and skilled workforces in culturally appropriate, place-based governance structures.

Rod Kennett *Australian Institute for Aboriginal and Torres Strait Islander Studies, Canberra, Australia.*
Finn Danielsen *NORDECO, Copenhagen, Denmark.*
Kirsten M. Silvius *Virginia Tech University, Blacksburg, Virginia, USA.*
rod.kennett@aiatsis.gov.au

Check the rejects for fame bias too

Omid Mahian and colleagues suggest that scientific celebrity does not seem to influence the editors' choice of Correspondence items for publication (*Nature* **519,** 414; 2015), but their informal analysis has caveats beyond those they mention.

Famous authors may be more disposed to write letters to the editor than is the rest of *Nature*'s readership, for example. Any analysis for editorial bias should include data from rejected contributions as well (which were obviously unavailable to the authors). Then, applying a logistic regression or calculating an index of niche breadth, such as the proportional similarity index (P. Feinsinger *et al. Ecology* **62,** 27–32; 1981), would clinch the matter.

Perhaps a simpler demonstration of a lack of editorial bias might be the publication of this note, signed by a student who has no publishing record at all.

Davide Nespoli *University of Milan, Italy.*
davide.nespoli@unimi.it

ASTROPHYSICS

The slow death of red galaxies

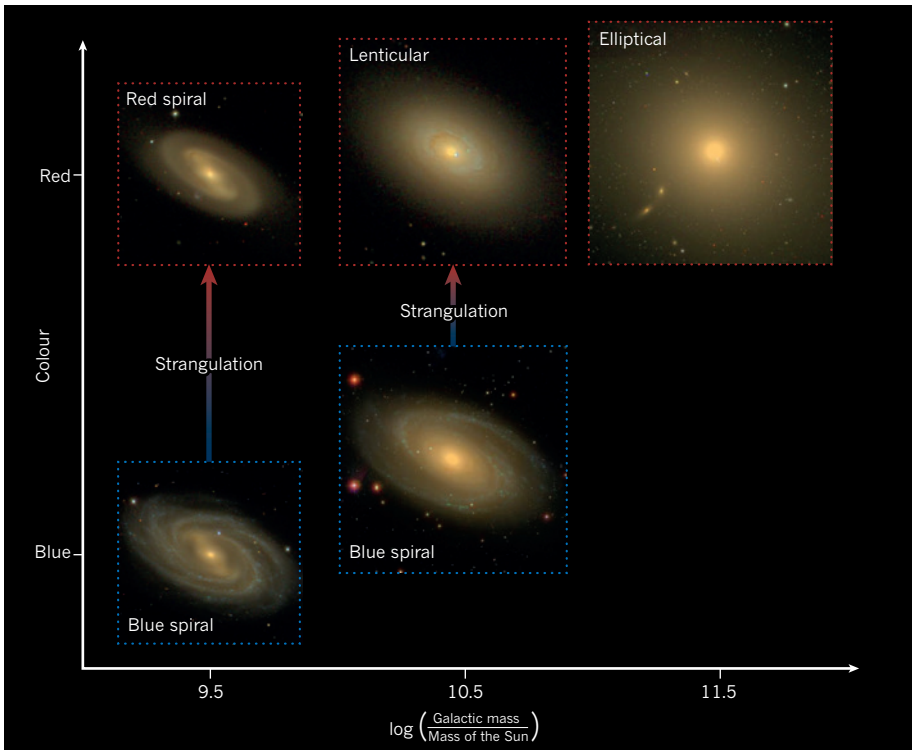
For most galaxies, the shutdown of star formation was a slow process that took 4 billion years. An analysis of some 27,000 galaxies suggests that 'strangulation' by their environment was the most likely cause. SEE LETTER P.192

ANDREA CATTANEO

In humans, death by strangulation is a slow process that takes about four minutes. During this time, the victim uses up oxygen in the lungs but keeps producing carbon dioxide, which remains trapped in the body. On page 192 of this issue, Peng *et al.*[1] present evidence of an analogous slow 'strangulation' process that ends the formation of stars in many galaxies by disrupting the supply of gas that accretes onto those galaxies from the environment. Instead of building up CO_2, the strangled galaxies accumulate metals — elements heavier than helium — produced by massive stars.

Almost 90 years ago, Edwin Hubble classified galaxies into three morphological types: spirals, ellipticals and lenticulars (Fig. 1). In spirals, stars form a disk and turn around in circles like the horses on a cosmic merry-go-round. Ellipticals are the wrecks of galaxy crashes, in which stars move chaotically in all directions. Lenticulars form an intermediate type between the two. Most spirals are blue because they contain young blue stars, but elliptical and lenticular galaxies contain little or no cold gas to make new stars, and so only old red stars remain. Historically, astronomers have been more interested in the morphologies of galaxies than in their colour. Attention switched to colour after the Sloan Digital Sky Survey (SDSS) measured the spectra of hundreds of thousands of galaxies.

The SDSS demonstrated that blue (star-forming) and red (passive) galaxies form distinct populations[2]. Since then, various hypotheses have been put forward to explain what causes galaxies to transition from one type to the other. Most revolve around two ideas. The first is that the gas in ellipticals and their surroundings is too hot to make stars and does not cool efficiently. The second is that the gas that could cool and make stars is kept hot or blown away by phenomena linked to the growth of supermassive black holes, which are found at the centres of all ellipticals[3]. The main motivation for considering violent expulsion scenarios comes from computer simulations of the formation of ellipticals in galaxy mergers. These simulations need a mechanism that gets rid of gas to avoid forming blue cores[4], which are rare in real ellipticals.



IMAGES: A. CATTANEO; SOURCE DATA: SDSS

Figure 1 | Colours and masses of galaxies. Blue galaxies produce new stars, whereas red galaxies do not. Spiral galaxies generally have relatively low masses (usually lower than $10^{10.5}$ times the mass of the Sun). Massive lenticular or elliptical red galaxies have masses approximately $10^{10.5}$–$10^{11.5}$ times the mass of the Sun. Any process that quenches star formation will shift galaxies from the blue to the red population. Peng *et al.*[1] present evidence for a slow process, called strangulation, that shuts down star formation and modifies the chemical composition of galaxies, but preserves their mass and structure.

Peng *et al.* present evidence that the formation of stars in most passive galaxies ended through a slow strangulation process. The authors compared the metal content of the stars in approximately 23,000 passive galaxies from the SDSS with that of a control sample of about 4,000 star-forming galaxies, also from the SDSS. They discovered that the metal content of the former is systematically larger than that of the latter, at least for galaxies that have stellar masses up to 100 billion times the mass of the Sun, the limit mass (M_*) above which galaxies become scarce. This constitutes evidence for galactic 'suffocation', in the same way that high levels of CO_2 in the blood of a corpse suggest suffocation.

From the difference in the metal content of the stars of passive and star-forming galaxies, Peng *et al.* inferred a delay of 4 billion years (or 2 billion years for galaxies close to M_*) between the time that gas stopped being supplied and the time that star formation ended. This delay is consistent with the mean age difference between passive and star-forming galaxies (about 4 billion years at all masses).

As any forensic scientist will tell you, suffocation does not imply strangulation. But the difference in metal content is higher for galaxies in groups and clusters than it is for isolated galaxies, suggesting that crowded environments strangle galaxies by disrupting the accretion of gas onto them. This disruption might occur either through ram pressure (the pressure exerted on a body moving through a fluid medium) or through tidal forces.

The slow shutdown of star formation inferred by Peng *et al.* from observations of galaxies with masses lower or equal to M_* contrasts with the fast shutdown behaviour of much larger galaxies, such as giant ellipticals. The stars of giant ellipticals have a low iron content because they were made on a short time span (less than 0.3 billion years for a galaxy of mass greater than $3\,M_*$)[5] — that is, there was not enough time for many type Ia supernova explosions, the source of iron. Because there are many more red galaxies below M_* than there are above, Peng *et al.* are correct to argue that strangulation is the main mechanism for star-formation shutdown. But the different chemical properties of low- and high-mass red galaxies imply that they must have formed through different routes. Morphological differences back this interpretation: red galaxies with masses above $M_*/3$ are all ellipticals or lenticulars; but below $M_*/3$, 40% of them are red spirals[6], as would be expected if strangulation has occurred.

There is also a difference between ellipticals and lenticulars. Most elliptical galaxies that have stellar masses in the range of $1\,M_*$ to $2\,M_*$ were already in place at a redshift of between 2 and 3, a period when the Universe was about one-fifth of its present age. Lenticulars appeared more gradually, replacing a pre-existing population of spiral and irregular galaxies (M. Huertas-Company, personal communication). The ratio of ellipticals to lenticulars is approximately 3:5 for galaxies of approximately $1.5\,M_*$, but lenticulars should predominate in the mass range explored by Peng and colleagues. Further evidence for the presence of two evolutionary tracks has been obtained by comparing the evolution of the star-formation rate of galaxies with the evolution of the galaxies' size[7].

Cosmological models[8] of the formation and evolution of galaxies predict two mechanisms by which star formation shuts down. In these models, galaxies more massive than $1.5\,M_*$ reside at the centres of groups and clusters. They grow extremely rapidly until, at a redshift of about 3 (which corresponds to when the Universe was one-quarter of its current size), they attain the critical mass at which infalling gas is effectively shock-heated. Some violent phenomenon then quenches star formation. However, the models also predict that galaxies below $0.6\,M_*$ become red much later (at a redshift of less than 0.5) and much more gradually, in most cases because they stop accreting gas after becoming part of a group or a cluster. Thanks to Peng and colleagues' work, this second theoretical prediction is now an observational fact. ■

Andrea Cattaneo *is in the Observatoire de Paris, GEPI, 75014 Paris, France.*
e-mail: andrea.cattaneo@obspm.fr

1. Peng, Y., Maiolino, R. & Cochrane, R. *Nature* **521,** 192–195 (2015).
2. Baldry, I. K. *et al. Astrophys. J.* **600,** 681–694 (2004).
3. Cattaneo, A. *et al. Nature* **460,** 213–219 (2009).
4. Di Matteo, T., Springel, V. & Hernquist, L. *Nature* **433,** 604–607 (2005).
5. Thomas, D., Maraston, C., Bender, R. & Mendes de Oliveira, C. *Astrophys. J.* **621,** 673–694 (2005).
6. Bell, E. F., McIntosh, D. H., Katz, N. & Weinberg, M. D. *Astrophys. J. Suppl.* **149,** 289–312 (2003).
7. Barro, G. *et al. Astrophys. J.* **765,** 104 (2013).
8. Cattaneo, A., Woo, J., Dekel, A. & Faber, S. M. *Mon. Not. R. Astron. Soc.* **430,** 686–698 (2013).

NEUROSCIENCE

Internal compass puts flies in their place

An analysis reveals that fruit-fly neurons orient flies relative to cues in the insects' environment, providing evidence that the fly's brain contains a key component for drawing a cognitive map of the insect's surroundings. SEE ARTICLE P.186

THOMAS R. CLANDININ & LISA M. GIOCOMO

Animals need accurate navigational skills as they go about their everyday lives. Many species, from ants to rodents, navigate on the basis of visual landmarks, and this is complemented by path integration, in which neuronal cues about the animal's own motion are used to track its location relative to a starting point. In mammals, these different types of navigation are integrated by neurons called head-direction cells[1]. In this issue, Seelig and Jayaraman[2] (page 186) provide the first evidence that certain neurons in fruit flies have similar properties to head-direction cells, encoding information that orients the insects relative to local landmarks.

Head-direction cells act as a neuronal compass that generates a cognitive map of an animal's environment. The activity of each head-direction cell increases as the animal faces a particular direction, with different cells preferentially responding to different directions[1,3]. Rather than certain cells always responding to north, south and so on, the direction in which the cells fire is set up arbitrarily when the animal encounters new visual landmarks. The signals are then updated by self-motion cues as the animal navigates. Studying head-direction cells in mammals is challenging because of the complexity of the mammalian brain. By contrast, the small fly brain is a good model for studying neuronal activity.

Seelig and Jayaraman recorded the activity of a specific population of neurons in the fruit-fly central complex, a region that spans the midline of the brain and which coordinates movement with visual cues. The authors

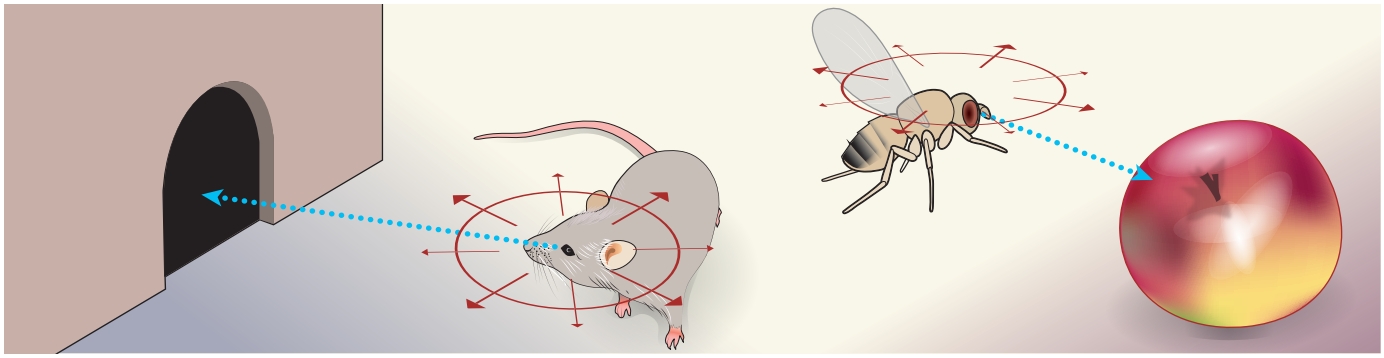


Figure 1 | A shared view of navigation. Mammals navigate on the basis of a combination of visual (blue dotted arrow) and self-motion cues. These cues are integrated by head-direction neurons to generate an internal compass (depicted in red) that helps the animal to orient itself in its surroundings. Seelig and Jayaraman[2] report that the neurons of fruit flies also process visual and self-motion cues to enable navigation, suggesting that flies, too, may have a cognitive compass.

placed flies on a ball in a virtual-reality arena of visual displays, and fixed the insects' heads in place. The flies explored the arena by walking on the ball, and the authors monitored turning behaviour (calculated from the ball's movements) and neuronal activity. Neurons in the central complex showed highly tuned responses that encoded the fly's orientation relative to a visual cue from the arena. The researchers showed that such orientation responses are generated by a population of highly active neurons, which together form a single 'bump' of neuronal activity that marks the direction in which the fly is facing relative to the visual cue. This activity bump seamlessly rotates around a ring of neurons in concert with the fly's location relative to the cue, with changes in the cue's location eliciting a concordant shift in the position of the activity bump.

Do the orientation responses encoded by fly neurons reflect the convergence of self-motion and landmark cues, as in head-direction cells? In support of this theory, increasing the complexity of the visual cue did not affect the orientation response. This indicates that neuronal activity predominantly reflects the fly's location relative to the visual landmark, rather than being a specific feature of the cue itself.

Under normal circumstances, animals integrate both self-motion and landmark cues. However, if one type of cue is missing, the animal must rely on the other. Seelig and Jayaraman allowed flies to explore the virtual-reality arena in the dark, thus eliminating landmark cues. Orientation responses remained stable, indicating that the response can be driven by self-motion, but they drifted slowly over time, showing that the accuracy of the response depends on both types of cue. Moreover, the fact that flies preserve a neuronal memory of orientation even in the absence of the visual cue that created it demonstrates that their navigational behaviour is much more than a simple sensorimotor reflex.

Thus, the current study points to many parallels between the orientation responses in flies and mammalian head-direction signals[1]. Both show increased activity when the animal faces a particular direction, irrespective of its location at the time. Both are maintained by internal self-motion cues, with familiar visual cues establishing each cell's directionality[3]. And, as in flies, orientation in rodents remains coherent in the dark, but drifts slowly over time[4,5]. These similarities demonstrate the sophistication of the insect navigational system. The mammalian system is thought to help generate an internal representation of space by acting as a neuronal compass (Fig. 1)[6]. The possibility that the fly brain contains at least one of the components needed to create a similar cognitive map is intriguing.

This study also improves our understanding of the mechanisms underlying orientation responses in both invertebrates and vertebrates. Head-direction cells are hypothesized[5] to arise from networks of neurons, dubbed 'attractor networks'. These networks are thought to adopt a ring-like architecture that allows a single activity bump to move around the network in concert with an animal's movement. Although some evidence consistent with this hypothesis has been observed in the past year in small numbers of rodent neurons[7,8], large-scale dynamics remain difficult to observe in mammalian neuronal networks. Seelig and Jayaraman observe classic traits of a ring-like attractor network in flies, including coherent activity in the absence of visual inputs, and slow error accumulation[9] — some of the most direct evidence to date for attractor networks as the mechanism that generates orientation responses. The possibility that ring-like attractor networks are evolutionarily conserved raises the exciting prospect that similar internal computational principles are used to calculate orientation in disparate species.

However, the role of self-motion compared to translational movement — movement of an animal's body position through the environment relative to the cue — in driving orientation responses remains a mystery. In rodents, head-direction responses remain coherent despite translational movement. For example, if a head-direction cell becomes established as north-preferring in response to a visual cue to the northeast, it will remain tuned to the north even if the animal moves such that the visual cue is now to its southeast. The flies in Seelig and Jayaraman's preparation, however, do not experience this type of translational movement, because the fly's body position relative to the cue remains fixed owing to the design of the virtual-reality system. It will be interesting to determine how fly orientation neurons respond to translation.

The authors' work provides insight into the neuronal basis of navigation. It was already known that insects use path integration[10], use polarized light to orient relative to the Sun[11] and show behaviours indicative of spatial memory[12]. But Seelig and Jayaraman provide the first evidence for one of the components required to construct a cognitive map — a computation previously thought the preserve only of vertebrates. This finding paves the way for dissecting the neuronal basis of navigation by leveraging the powerful genetic tools and tractable neuronal circuits available in flies. ■

Thomas R. Clandinin *and* **Lisa M. Giocomo**
are in the Department of Neurobiology, Stanford University, Stanford, California 94305, USA.
e-mail: trclandinin@gmail.com; giocomo@gmail.com

1. Taube, J. S., Muller, R. U. & Ranck, J. B. Jr *J. Neurosci.* **10,** 420–435 (1990).
2. Seelig, J. D. & Jayaraman, V. *Nature* **521,** 186–191 (2015).
3. Taube, J. S., Muller, R. U. & Ranck, J. B. Jr *J. Neurosci.* **10,** 436–447 (1990).
4. Goodridge, J. P., Dudchenko, P. A., Worboys, K. A., Golob, E. J. & Taube, J. S. *Behav. Neurosci.* **112,** 749–761 (1998).
5. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. *Adv. Neural Inf. Process. Syst.* **7,** 173–180 (1994).
6. Taube, J. S. *Annu. Rev. Neurosci.* **30,** 181–207 (2007).
7. Bjerknes, T. L., Moser, E. I. & Moser, M.-B. *Neuron* **82,** 71–78 (2014).
8. Peyrache, A., Lacroix, M. M., Petersen, P. C. & Buzsáki, G. *Nature Neurosci.* **18,** 569–575 (2015).
9. Zhang, K. *J. Neurosci.* **16,** 2112–2126 (1996).
10. Wehner, R. & Srinivasan, M. V. *J. Comp. Physiol.* **142,** 315–338 (1981).
11. Heinze, S. & Homberg, U. *Science* **315,** 995–997 (2007).
12. Ofstad, T. A., Zuker, C. S. & Reiser, M. B. *Nature* **474,** 204–207 (2011).

MICROSCOPY

Quantum control of free electrons

Optical pulses have previously been used to place the electrons in the beam of an electron microscope into well-defined energy states. These electrons can now be put in a quantum superposition of those states. SEE LETTER P.200

MATHIEU KOCIAK

Quantum mechanics is a daily affair for electron microscopists. The researchers routinely focus their microscopes' electron beams on samples to create quantum interference patterns that reveal information about the samples' atomic and molecular structure. In an exciting paper in this issue, Feist *et al.*[1] (page 200) demonstrate how they have used light focused on a nanostructure to make the fast electrons of one such electron beam exhibit another type of quantum behaviour — Rabi oscillations.

Quantum systems have properties that their classical counterparts lack. For two-state

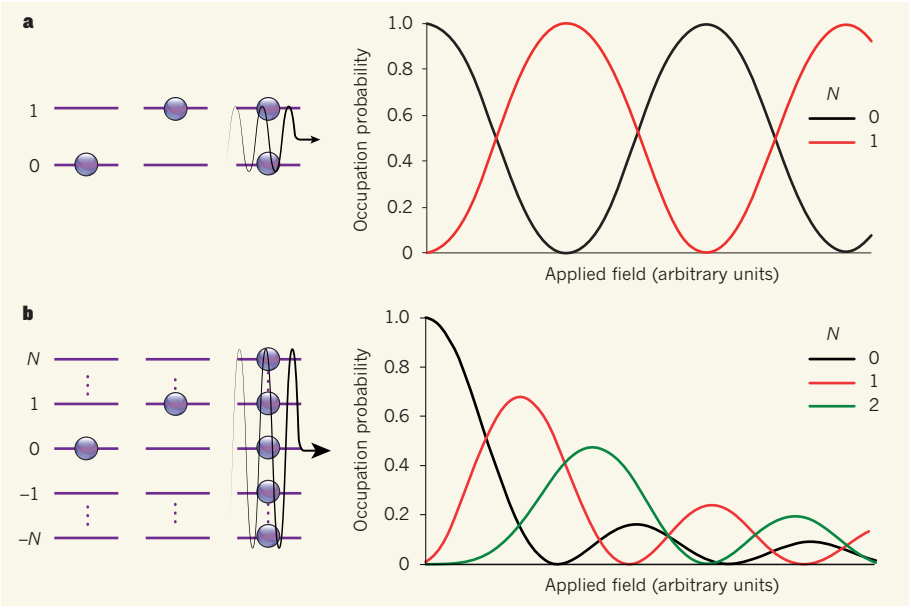


Figure 1 | Quantum superpositions. **a**, In a two-level system, an electron can be in either of two stationary states (0 or 1). An applied strong, oscillating electromagnetic field (black line) can put the electron into a coherent superposition of states. The probabilities of detecting the electron in either of the states oscillate with increasing strength of the field (Rabi oscillations). **b**, The freely propagating electrons in the electron beam of an electron microscope, such as that studied by Feist *et al.*[1], can be placed in a large set of stationary states using an appropriately shaped electromagnetic field: state 0, which is determined by the nominal accelerating voltage of the microscope; states −1 to $-N$, which have the energy of state 0 minus integer multiples of the electromagnetic field's energy; and states 1 to $+N$, which have the energy of state 0 plus integer multiples of the field's energy. For a strong field, each electron is put in a superposition of these states, and the probability that the electron is in any of these states oscillates with the field strength (unconventional, multilevel Rabi oscillations). For simplicity, only a subset of the oscillations is shown. (Plot in **b** adapted from ref. 1.)

quantum systems, Rabi oscillations between the two, otherwise stationary, states of the system are one such property. To understand these oscillations, consider an electron in an atom that has two energy states, a ground state and an excited state (Fig. 1a). The electron will stay in one of the states forever if no external perturbation is applied to it. But if, for example, an electromagnetic field is applied to the system, the probability of finding the electron in either of the states will oscillate with time and the strength of the electromagnetic field: these are the Rabi oscillations. The oscillation frequency, called the Rabi frequency, depends mainly on the strength of the perturbation. Stopping the oscillations at any time leaves the electron in a coherent quantum superposition; that is, in a combination of both stationary states at the same time. Rabi oscillations are therefore commonly used to prepare quantum systems in a superposition of states. Such superposition provides a means of encoding information in quantum-information technologies.

In the electron microscope used by Feist *et al.*, the freely propagating electrons of the electron beam can be placed in a large set of stationary states[2] (Fig. 1b). If Rabi oscillations between these states are to be observed, the probability of the electrons being in a given energy state must be measured. This probability measurement can be easily done using an electron energy-loss spectrometer (EELS), a device that can be added to an electron microscope. It is usually used to produce a spectrum of the electron intensity that is transmitted through the sample under study as a function of the energy loss caused by scattering from the sample. The EELS spectrum of freely propagating electrons consists of peaks centred at the energy of the electrons, and measuring the height of the peaks gives the probability of finding them in these energy states.

However, putting freely propagating electrons in a coherent quantum superposition is much more difficult than measuring their probabilities of being in a given state. Unlike an electron bound to an atom, a fast electron moving in free space will generally not couple to an electromagnetic field such as an optical plane wave. But it is well established that such an electron can easily couple to an evanescent wave — a wave that does not propagate but that decays exponentially with distance from the boundaries of objects, such as nanostructures, at which they are formed. Electron microscopy is regularly used for imaging optical excitations in nanostructures at high spatial resolution through the nanostructures' evanescent fields[3].

By using synchronized femtosecond (1 femtosecond is 10^{-15} seconds) pulses of electrons and photons in an electron-microscope set-up, researchers have previously coupled photons to fast electrons through the evanescent light field of a nanostructure (an individual carbon nanotube or a silver nanowire)[2]. They showed that, in such cases, electrons could absorb or emit photons many times, resulting in spectra of electron-energy loss (or gain) consisting of a series of peaks evenly spaced according to the energy of the photons. The peaks' intensity, and thus the probability of finding an electron in a given state, decreased monotonically with the energy of the electron loss (or gain). Soon after this remarkable experiment, others predicted[4] that, using a similar experimental set-up, quantum superpositions of freely propagating electron states should be observable for longer-lasting photon pulses as a series of peaks with oscillating energies of electron loss (or gain).

And this is exactly what Feist *et al.* have demonstrated experimentally in their study. The authors observed a quantum superposition of freely propagating electron states by focusing femtosecond electron pulses on a sharp, nanometre-sized gold tip illuminated by picosecond laser pulses (1 picosecond is 10^{-12} s). The tip turned the laser's plane wave into an evanescent one, allowing strong coupling of the large number of freely propagating electron states. Each state's probability, encoded in the intensity of each peak in an EELS spectrum, oscillates at its own pace, leading to unconventional, multilevel Rabi oscillations (Fig. 1b).

Feist and colleagues' experimental achievement lies largely in the development of an electron gun that generates electron pulses with high brightness. Such an electron gun allows a relatively narrow electron beam (one with a diameter of about 15 nanometres) to be formed. A wider beam would average the evanescent field in such a manner that the Rabi oscillations and the coherent quantum superposition would be lost. By experimentally introducing the field of free-electron quantum optics, this work projects free electrons into the world of quantum information — although it will be technologically demanding to transform an electron microscope into a quantum-information processor. However, as Feist *et al.* also demonstrated, the control of the beam's freely propagating electrons leads to effects other than quantum superposition, including the formation of a train of attosecond (one billion-billionth of a second) electron pulses. Such pulses could find applications in ultrafast electron spectroscopy and microscopy. ■

Mathieu Kociak *is at the Laboratoire de Physique des Solides, CNRS/Université Paris Sud, 91400 Orsay, France.*
e-mail: mathieu.kociak@u-psud.fr

1. Feist, A. *et al. Nature* **521,** 200–203 (2015).
2. Barwick, B., Flannigan, D. J. & Zewail, A. H. *Nature* **462,** 902–906 (2009).
3. García de Abajo, F. J. *Rev. Mod. Phys.* **82,** 209–275 (2010).
4. García de Abajo, F. J., Asenjo-Garcia, A. & Kociak, M. *Nano Lett.* **10,** 1859–1863 (2010).

MICROBIOLOGY

Fungus against the wall

A compound derived from plant cell-wall material that is a waste product of biofuel manufacture has been found to have fungicidal properties: it interacts with a carbohydrate called β1,3 glucan, thus compromising the integrity of fungal cells.

PAUL O'MAILLE

Fungal infections are a major problem in agriculture, and as the world's population grows, a rising tide of resistance to antifungal agents is threatening global food security. This threat has led to an intensification in the use of fungicidal agrochemicals, which may themselves put human health at risk and damage the environment. New fungicidal agents are therefore desperately needed; and now it seems that help is on the way, but from unlikely sources. Writing in *Proceedings of the National Academy of Sciences*, Piotrowski and colleagues[1] report the discovery of molecules in the waste products of biofuel manufacture that possess antifungal activity against plant pathogens.

Dwindling oil reserves have spurred the development of technologies that use lignocellulosics — dry plant matter such as grass — as a renewable source of sugar to produce biofuels and bio-based chemicals. A key step in the process is the hydrolysis of polysaccharides in the plant cell wall, which liberates sugar monomers so they can be fermented by yeast[2] (Fig. 1a). But other molecules are also liberated by hydrolysis, including diferulates, metabolites of the phenylpropanoid family of natural products. These molecules crosslink polysaccharides to each other, and to lignin polymers, to stiffen the plant cell wall, but the waste products stunt the growth of the yeast, inhibiting the fermentation process. Piotrowski and colleagues saw this problem as an opportunity for 'bioprospecting', reasoning that these molecules might be an untapped resource of antifungal agents.

The authors screened diferulates for antifungal activity by testing the molecules' ability to inhibit growth of a fungus, the budding yeast *Saccharomyces cerevisiae*. They identified two active compounds that inhibited growth, the most potent of which they dubbed poacic acid because of its prevalence in grasses of the Poaceae plant family. Importantly, poacic acid had comparable antifungal activity to two widely used agricultural fungicides.

Next, Piotrowski *et al.* investigated how poacic acid exerts its effect, by challenging a pooled mixture of approximately 4,000 yeast strains with the compound. Each strain carried a different gene deletion that was marked by a unique molecular barcode — a short sequence of DNA that does not affect the cell's function but can be used to identify the strain. After exposing the cells to poacic acid, the authors took a head count with next-generation DNA sequencing, measuring the frequency of each barcode in the population to determine the relative sensitivity or resistance of each strain to the fungicide[3]. The result was clear: deletion of genes involved in cell-wall organization conferred poacic-acid sensitivity, suggesting that the compound affects the yeast cell wall, thus mirroring the mechanism of many known antifungal compounds[4].

The authors then tested poacic acid in combination with two antifungal compounds that compromise cellular integrity[5] and that are normally used to treat human fungal infections: fluconazole, which targets the biosynthesis of ergosterol, an essential component of the fungal cell membrane; and capsofungin, which inhibits[6] the biosynthesis of β1,3 glucan, a polysaccharide that acts as both a barrier and structural support in the fungal cell wall. Perhaps not unexpectedly, poacic acid acted in synergy with these compounds, indicating that each drug strikes the same or related pathways, albeit through different targets[7].

To further investigate how poacic acid alters the cell wall, Piotrowski *et al.* used high-dimensional microscopy to study yeast cells treated with poacic acid. The neck between the yeast cell and the daughter that buds off it was unusually wide in treated cells, and the degree of morphological variability in the cell population was higher than normal — similar changes to those caused by other drugs that disrupt the cell wall[8] (Fig. 1b). Poacic acid also caused rapid influx of propidium iodide dye into the cell, consistent with cell-wall disruption. In a stroke of resourcefulness, the authors used the natural fluorescence of poacic acid to study its location within the cell. The cell's surface lit up under the microscope, suggesting that poacic acid interacts directly with the β1,3 glucan layer. The researchers then incubated poacic acid with purified fungal β1,3 glucan — the glucan particles began to fluoresce, confirming the interaction.

Given these promising results, the researchers next tested the potency of poacic acid against key agricultural fungal pathogens. Preliminary studies showed that poacic acid inhibited the fungi *Sclerotinia sclerotiorum* and *Alternaria solani* in a dose-dependent manner and protected soya-bean leaves from the oomycete (fungus-like) pathogen *Phytophthora sojae*, suggesting that poacic

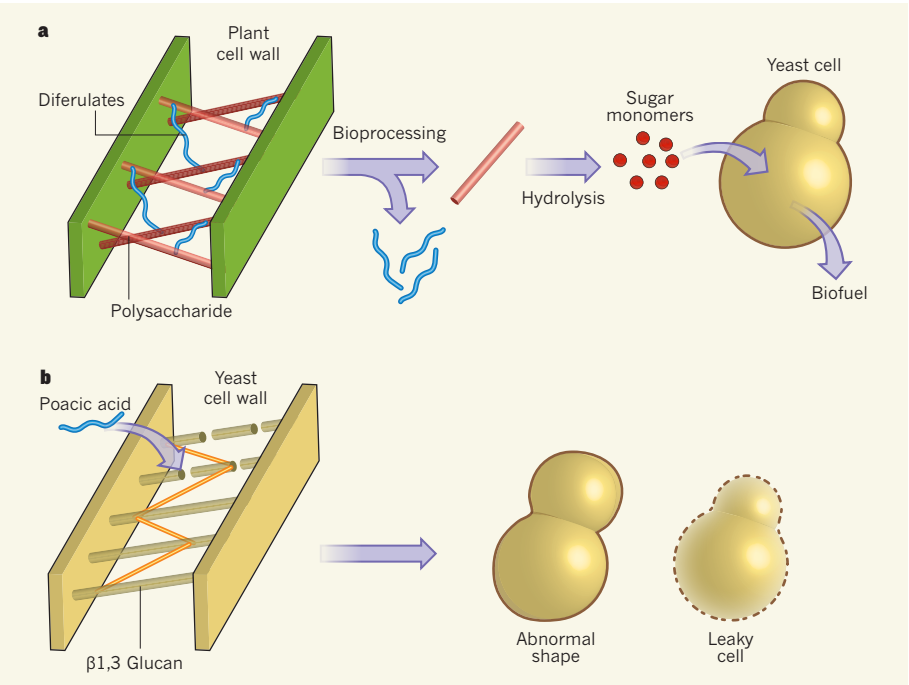


Figure 1 | Waste of a fungicide. **a**, Grasses of the Poaceae family have cell walls fortified with polysaccharides that are crosslinked with metabolites called diferulates, which act as an anchor for the formation of lignin polymers (not shown). Bioprocessing of such grasses liberates the polysaccharides for hydrolysis to sugar monomers, from which yeast can produce biofuels and bio-based chemicals. The process also liberates diferulates such as poacic acid as waste products. **b**, Piotrowski *et al.*[1] report that poacic acid has antifungal activity against yeast cells. The compound attacks the carbohydrate β1,3 glucan, which acts as a fortification in the yeast cell wall. As a result, cells become leaky and adopt abnormal shapes.

acid has broad-spectrum activity.

Although the ability of poacic acid to target β1,3 glucan and to disrupt cell walls in yeast is clear, more work is needed to establish the extent to which these effects are mirrored in agricultural pathogens and to gain deeper insights into the biochemical mechanism of action of poacic acid. Metabolomics (the study of metabolites within an organism, cell or tissue) provides a powerful tool with which to examine specific changes in the cellular biochemicals induced by drug action, and it has been successfully used to study pesticides[9]. When applied to poacic acid and other antifungals, metabolomics may help us to define the specific enzymatic targets that underlie fungicidal action. This could lead to the discovery of synergistic drug pairings and might help to explain why some fungicides are toxic to humans[10].

Piotrowski and colleagues' demonstration of synergistic effects between poacic acid and human antifungal drugs is promising, suggesting that poacic acid and derivatives might have roles in medicine. But although poacic acid is probably also synergistic with agricultural fungicides, this remains to be examined and will be a crucial part of future studies. Finally, before poacic acid can be translated into a market-ready fungicide there are still major hurdles to be overcome, including testing its effectiveness in field trials, efficacy against more-diverse pathogens and persistence in the environment, and establishing whether it is safe for humans.

Although much work is still needed, this study is a timely illustration of how the right biological techniques, combined with resourcefulness — in this case discovering fungicides in waste biomass — can be used to gain insights into drug action. The work also shows how understanding drug mechanisms can help to identify synergistic drug pairings. This strategy holds great potential for rationally designed combination therapies that could extend the potency and lifetimes of existing fungicides. On its own, in combination with other fungicides or as a lead compound for further development, poacic acid is a potentially valuable new weapon in our armamentarium of antifungals. ■

Paul O'Maille *is at the John Innes Centre and Institute of Food Research, Norwich NR4 7UH, UK.*
e-mail: paul.o'maille@jic.ac.uk

1. Piotrowski, J. S. *et al. Proc. Natl Acad. Sci. USA* **112,** E1490–E1497 (2015).
2. Sun, Y. & Cheng, J. *Bioresour. Technol.* **83,** 1–11 (2002).
3. Parsons, A. B. *et al. Cell* **126,** 611–625 (2006).
4. Latgé, J.-P. *Mol. Microbiol.* **66,** 279–290 (2007).
5. Kiraz, N. *et al. Antimicrob. Agents Chemother.* **54,** 2244–2247 (2010).
6. Kurtz, M. B. & Douglas, C. M. *J. Med. Vet. Mycol.* **35,** 79–86 (1997).
7. Cokol, M. *et al. Mol. Syst. Biol.* **7,** 544 (2011).
8. Okada, H., Ohnuki, S., Roncero, C., Konopka, J. B. & Ohya, Y. *Mol. Biol. Cell* **25,** 222–233 (2014).
9. Aliferis, K. A. & Jabaji, S. *Pestic. Biochem. Physiol.* **100,** 105–117 (2011).
10. Bouhifd, M., Hartung, T., Hogberg, H. T., Kleensang, A. & Zhao, L. *J. Appl. Toxicol.* **33,** 1365–1383 (2013).

EVOLUTION

Steps on the road to eukaryotes

A new archaeal phylum represents the closest known relatives of eukaryotes, the group encompassing all organisms that have nucleated cells. The discovery holds promise for a better understanding of eukaryotic origins. SEE ARTICLE P.173

T. MARTIN EMBLEY & TOM A. WILLIAMS

There are many competing hypotheses[1] for how eukaryotic cells, which contain a nucleus and other membrane-bound organelles, evolved from their prokaryotic ancestors, whose cells lack a nucleus. But testing these theories has been difficult owing to a lack of known intermediate stages in the prokaryote-to-eukaryote transition. On page 173 of this issue, Spang *et al.*[2] describe a prokaryotic lineage that is more closely related to eukaryotes than any yet sampled and that shares with eukaryotes several genes previously thought to define aspects of eukaryotic biology. This technically outstanding paper has far-reaching implications for how we view early eukaryotic evolution, including our own deep ancestry.

In most textbooks the cellular world is divided into three domains[3]: the eukaryotes (Eukarya) and two distinct prokaryotic groups, the Bacteria and the Archaea. In the classical three-domains tree, the eukaryotes are separated from a common prokaryotic ancestor that they share with Archaea by a long branch that has been variously interpreted as representing a long period of time with unsampled diversity, a high rate of evolution in the eukaryotic ancestor, or the extinction of intermediate forms. Surveys of environmental microbial diversity using the tools of molecular biology have sought to populate this long branch, but have so far failed to identify any fundamentally new eukaryotic groups. In the three-domains tree, the eukaryotes appear fully formed, with almost all of the cellular complexity that we associate with modern eukaryotes already in place[1].

The three-domains tree is the most visible image depicting the diversity of cellular life, but it has not gone unchallenged. An alternative two-domains tree, in which the eukaryotic lineage originated within the archaeal domain, has gathered support from recent phylogenetic analyses[4–6] and is now arguably the favoured hypothesis. In this tree, the eukaryotes are related to a diverse group of Archaea called the TACK superphylum[7]. Thus, unlike the three-domains tree, the two-domains tree includes an explicit prediction about where we should look for closer relatives of the eukaryotic

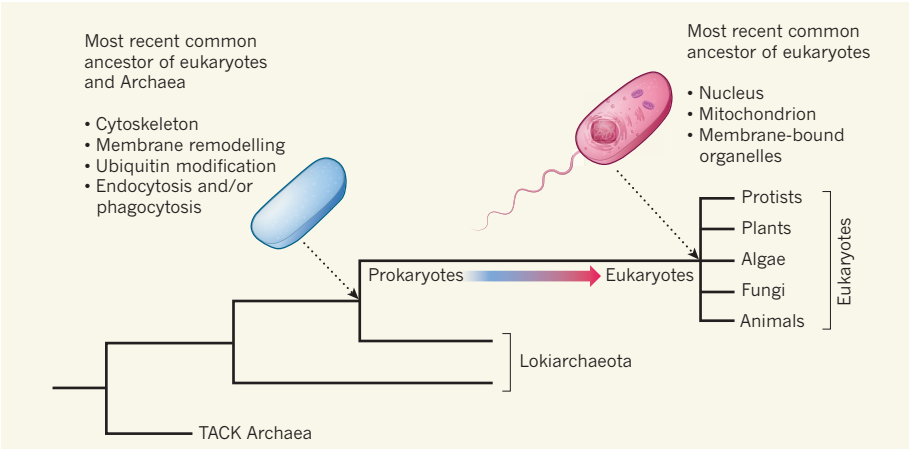


Figure 1 | Lokiarchaeota are the closest known prokaryotic relatives of eukaryotes. Phylogenetic trees presented by Spang *et al.*[2] place eukaryotes within the Lokiarchaeota — a new group of Archaea described by the authors. The genomes of these Archaea contain more eukaryotic-like genes than other known Archaea. This finding implies that some of the defining features of eukaryotes — including a cytoskeleton, membrane remodelling, ubiquitin modification and the capacity for endocytosis and/or phagocytosis — might have already evolved in the last common ancestor of eukaryotes and Archaea. Spang and colleagues' findings suggest that it is likely that other new lineages will be found to further close the evolutionary gap between Archaea and eukaryotes, increasing the precision with which we can identify when key cellular innovations such as the nucleus, mitochondrion and endoplasmic reticulum first evolved.

ancestral lineage. Spang *et al.* report the first spectacular results of that search.

The authors sequenced the combined genomes (metagenomes) from samples of marine sediments that had been enriched for members of the Deep-Sea Archaeal Group, a lineage related to the TACK Archaea that has not been cultured in the laboratory. Using cutting-edge computational methods, Spang *et al.* reconstructed one largely complete and two partial genomes from closely related members of the Deep-Sea Archaeal Group, for which they propose the new name Lokiarchaeota. Using slowly evolving marker genes, the authors constructed phylogenetic trees that place eukaryotes within the Lokiarchaeota at the base of the TACK superphylum, suggesting that Lokiarchaeota are the closest prokaryotic relatives of eukaryotes yet discovered (Fig. 1).

Consistent with the relationship implied by the authors' trees, the Lokiarchaeum genome contains more eukaryotic signature genes than do other prokaryotes, including genes encoding actin proteins, components of a primordial vesicle-trafficking complex, a ubiquitin-modifier system and a diverse array of small GTPase enzymes belonging to the Ras superfamily. Making reliable evolutionary trees for ancient relationships is difficult[5], especially given that the split between eukaryotes and the Lokiarchaeota may have occurred more than 2 billion years ago. So it is this combination of enhanced 'eukaryote-like' genome content and the phylogeny that make the case for a close relationship between eukaryotes and Lokiarchaeota so convincing.

In the absence of cultured Lokiarchaeota, the cellular manifestation of this enhanced protein repertoire can be only indirectly inferred. Nevertheless, the genome hints at an organism that has a dynamic, actin-based cytoskeleton, vesicular-trafficking and membrane-remodelling capabilities, and the potential for uptake of materials from the environment by endocytosis and/or phagocytosis. These are all traits that could have enabled the common ancestor of all eukaryotes to engulf the bacterial symbiont that became the progenitor of the mitochondrion, a vital organelle of modern eukaryotes[1]. The presence of these genes in an otherwise unambiguously archaeal genome is consistent with hypotheses proposing that an archaeon was the host for that fundamental event in the development of eukaryotic cells[1,8].

The discovery of the Lokiarchaeota provides strong evidence that phylogenetics can be used to infer ancient relationships and, in combination with single-cell and metagenomic sequencing, provides a powerful toolkit for testing ideas about the origins of the component parts of eukaryotic cells[1,5,6]. The same techniques can be applied in future studies to pinpoint the source of the mitochondrial endosymbiont, and to investigate the origins of the many bacterial genes that comprise a major component of eukaryotic genomes. The sequences of genes encoding ribosomal RNA molecules, which are commonly used for taxonomic classification, can help to identify environmental samples enriched in Lokiarchaeota and its relatives. This will facilitate the isolation and cultivation of such cells for detailed study of their biology and metabolism.

The identification of Lokiarchaeota so early in the history of this nascent field suggests that more-closely related archaeal relatives of eukaryotes will soon be discovered. Some of these may be related to the Deep-Sea Archaeal Group[9], an enormously diverse and abundant radiation of Archaea for which the Lokiarchaeota genomes are the first available. The genomes and cellular features of these relatives may provide a more detailed picture of the most recent common ancestor of eukaryotes and Archaea, and may help to resolve the timing of the innovations that are used to define eukaryotes. Not least, the discovery of Lokiarchaeum, and the promise of further discoveries from the vast, unexplored world of prokaryotic diversity, raises the tantalizing prospect that the investigation of eukaryotic origins can now enter the realm of testable science. ■

T. Martin Embley *and* **Tom A. Williams**
are at the Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.
e-mails: martin.embley@ncl.ac.uk; tom.williams2@ncl.ac.uk

1. Embley, T. M. & Martin, W. *Nature* **440,** 623–630 (2006).
2. Spang, A. *et al. Nature* **521,** 173–179 (2015).
3. Woese, C. R., Kandler, O. & Wheelis, M. L. *Proc. Natl Acad. Sci. USA* **87,** 4576–4579 (1990).
4. Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. *Proc. Natl Acad. Sci. USA* **81,** 3786–3790 (1984).
5. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. *Nature* **504,** 231–236 (2013).
6. McInerney, J. O., O'Connell, M. J. & Pisani, D. *Nature Rev. Microbiol.* **12,** 449–455 (2014).
7. Guy, L. & Ettema, T. J. G. *Trends Microbiol.* **19,** 580–587 (2011).
8. Lane, N. & Martin, W. *Nature* **467,** 929–934 (2010).
9. Jørgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. *Front. Microbiol.* **4,** 299 (2013).

This article was published online on 6 May 2015.

MOLECULAR BIOLOGY

Rap and chirp about X inactivation

Two new techniques identify proteins that directly interact with a non–protein–coding RNA called *Xist* to mediate inactivation of one X chromosome in female mammals. SEE LETTER P.232

ANNA ROTH & SVEN DIEDERICHS

In mammals, females have two X chromosomes, whereas males have only one. To ensure that the expression of genes from the X chromosome is equal between the two sexes, one X chromosome, dubbed Xi, is inactivated in all female cells except eggs. X-chromosome inactivation (XCI) involves the long non-coding RNA (lncRNA) *Xist*, but only a few proteins that directly bind to *Xist* have been identified[1,2]. The development of two techniques for isolating RNA-bound proteins (one described on page 232 of this issue[3] and one published in *Cell*[4]) now sheds light on more *Xist*-interacting proteins that have a role in XCI.

In 1961, the Lyon hypothesis posited that XCI occurs through the repression of genes on the Xi during early embryonic development and that this state is maintained as cells divide[5]. Decades later, the identification of *Xist*[6], which is expressed exclusively from the Xi, provided an explanation for how such gene silencing arises. During XCI initiation, the Xi becomes coated in *Xist*, which establishes a repressive environment in which the proteins involved in transcription are displaced from the chromosome. Two protein complexes, Polycomb repressive complex 1 (PRC1) and PRC2 (ref. 7), are subsequently drawn to the Xi and modify the histone proteins around which DNA is packaged. These changes establish a stable state of transcriptional repression. Finally, spreading of *Xist* along the entire chromosome leads to silencing of genes across the Xi[8].

Despite progress in understanding XCI, many mechanistic details remain unclear because only a few of the partner proteins required for gene silencing by *Xist* have been found[1,2]. The key to progress lies in identifying other proteins that directly interact with *Xist*. As such, unbiased 'pull-down' techniques that capture the proteins bound to a specific RNA are much sought after.

McHugh *et al.*[3] and Chu *et al.*[4] present highly sensitive techniques — named RNA antisense purification–mass spectrometry (RAP–MS) and comprehensive identification of RNA-binding proteins by mass spectrometry (ChIRP–MS), respectively — that can

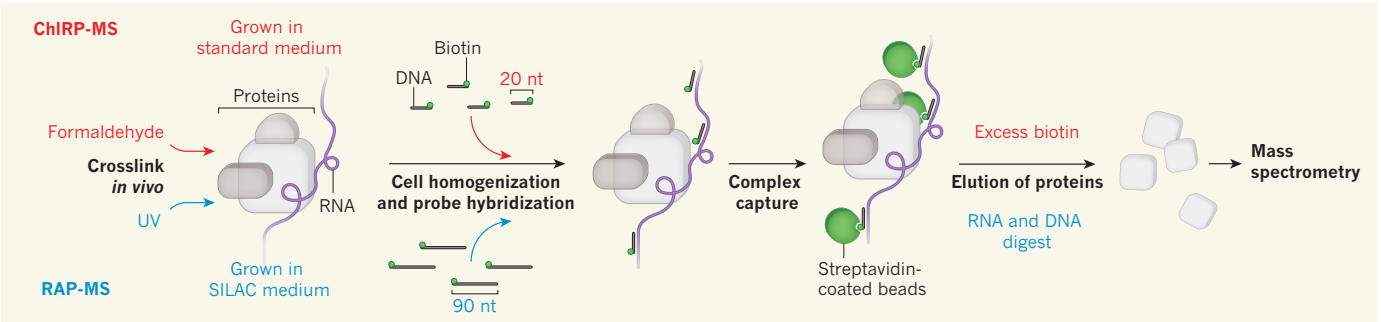


Figure 1 | Capturing RNA–protein complexes. Two groups[3,4] have developed techniques, called comprehensive identification of RNA binding proteins by mass spectrometry (ChIRP–MS) and RNA antisense purification–mass spectrometry (RAP–MS), for identifying the proteins that interact with specific RNAs *in vivo*. In RAP–MS, cells are grown in a protein-labelling solution called SILAC (stable isotope labelling by amino acids in culture) medium, which allows a quantitative comparison of the proteins that interact with the RNA of interest and with a control RNA. In both techniques, RNA–protein interactions are preserved by crosslinking, through either ultraviolet irradiation or formaldehyde treatment. Cells are homogenized, and the RNA–protein complexes within are bound by DNA probes — strands of 90 or 20 nucleotides (nt) that are complementary to the RNA of interest, tagged with a biotin molecule at one end. The complexes are then captured on beads coated in streptavidin protein, which interacts strongly with biotin. In RAP–MS, the proteins are eluted from the beads by degradation (digestion) of the RNA and DNA. By contrast, in ChIRP–MS addition of extra biotin molecules interferes with biotin–streptavidin binding to displace the complexes. Finally, proteins are identified by mass spectrometry.

be used to identify the proteins that interact with *Xist in vivo* (Fig. 1). The methods both involve a crosslinking step that preserves naturally occurring RNA–protein complexes, which are then captured using DNA sequences (90 nucleotides long in RAP–MS, and 20 nucleotides long in ChIRP–MS) that are complementary to, and so bind to, the RNA sequence of interest. The DNA is tagged at one end with a biotin molecule, which interacts with a protein called streptavidin on magnetic beads. Only RNA–protein complexes bound to the biotin-tagged DNA are captured by the beads. Finally, the proteins are eluted from the beads and identified through mass spectrometry.

A key difference between the two protocols is that in RAP–MS, cells are grown in SILAC (stable isotope labelling by amino acids in culture) medium, which contains the amino acids lysine and arginine, labelled with either heavy or light isotopes. Proteins are then labelled with heavy or light isotopes depending on the medium used, allowing a quantitative comparison of the proteins pulled down with *Xist* compared to those pulled down with a control RNA molecule. By contrast, ChIRP–MS uses a standard culture medium.

Using ChIRP–MS, Chu and colleagues captured 81 *Xist*-bound proteins in four different types of mouse embryonic stem cell — one type that has the potential to differentiate into all bodily cell types, and three that are preparing to differentiate into more-specialized cells. Although most proteins bound *Xist* in all four types of cell, 19 were not found in the most primitive cell type. The two sets of proteins might therefore reflect stepwise changes in *Xist* binding patterns that arise as embryonic stem cells begin to differentiate.

The 'repeat A region' is an evolutionarily conserved RNA sequence close to the 5′ end of *Xist* that is implicated in transcriptional silencing[9]. Chu *et al.* found that *Xist* molecules that lack the repeat A region cannot bind to three of the proteins that are pulled down by the full-length RNA. One such protein is the transcriptional repressor Spen (also known as Sharp). The authors found that Spen directly interacts with the repeat A region *in vivo* and *in vitro*. Furthermore, they demonstrate that Spen and two other proteins, HnrnpK and HnrnpU, are involved in gene silencing during XCI, and provide evidence to suggest that HnrnpK might mediate PRC2 recruitment to the Xi.

McHugh *et al.* used RAP–MS to identify ten proteins that directly bind to *Xist*, three of which — Spen, Lbr and HnrnpU — are necessary for silencing of genes on the Xi. The authors report that Spen represses transcription in the early stages of XCI, by binding to its co-repressor protein, Smrt (ref. 10), which activates Hdac3 (ref. 11). This enzyme removes acetyl groups from histones, thereby promoting transcriptional repression. The authors propose that *Xist*–PRC2 interactions are independent of the repeat A region but are indirectly mediated by the Spen–Smrt–Hdac3 complex through an as-yet-unknown mechanism.

The development of techniques for analysing RNA–protein interactions is instrumental to the success of both of the current studies. Impressively, nine out of ten proteins identified by RAP–MS are also pulled down by ChIRP–MS, demonstrating the specificity of both techniques. The difference in the number of *Xist*-interacting proteins identified might arise from differences in methodology, the cell models and control RNAs used or the thresholds used to determine whether an interaction is significant. Each group validated its method using cellular non-coding RNAs with well-characterized binding proteins, rather than lncRNAs. Future studies should test the efficiency of the techniques using extensively studied and abundant lncRNAs, such as MALAT1 (ref. 12), and, most importantly, using less-abundant lncRNAs, because most lncRNAs exist in low numbers in cells.

Both studies provide some mechanistic insight into XCI. Neither work captured components of PRC2, supporting a model in which *Xist*–PRC2 interactions are indirectly mediated by HnrnpK and Spen or by PRC1. Alternatively, different isoforms of *Xist* might interact with different proteins. Defining the roles of other interaction partners in *Xist*-mediated XCI will shed more light on the complex mechanisms underlying this process.

More than 10,000 lncRNAs have now been discovered — some of which have vital roles in health and disease[13]. However, the cellular functions and molecular mechanisms of most lncRNAs are unknown. ChIRP–MS and RAP–MS are powerful and versatile techniques that have the potential to boost RNA research by reliably identifying the proteins that interact with any specific RNA sequence *in vivo*. Notably, they may enable the identification of entire lncRNA-binding networks, including proteins, DNA and RNA — a key step towards elucidating the pathways in which lncRNAs function. ■

Anna Roth *and* **Sven Diederichs** *are in the RNA Biology and Cancer Division, German Cancer Research Center, and at the Institute of Pathology, University Hospital Heidelberg, 69120 Heidelberg, Germany.*
e-mail: s.diederichs@dkfz.de

1. Wutz, A. *Nature Rev. Genet.* **12,** 542–553 (2011).
2. Sarma, K. *et al. Cell* **159,** 869–883 (2014).
3. McHugh, C. A. *et al. Nature* **521,** 232–236 (2015).
4. Chu, C. *et al. Cell* **161,** 404–416 (2015).
5. Lyon, M. F. *Nature* **190,** 372–373 (1961).
6. Brown, C. J. *et al. Nature* **349,** 38–44 (1991).
7. Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J. & Lee, J. T. *Science* **322,** 750–756 (2008).
8. Simon, M. D. *et al. Nature* **504,** 465–469 (2013).
9. Wutz, A., Rasmussen, T. P. & Jaenisch, R. *Nature Genet.* **30,** 167–174 (2002).
10. Shi, Y. *et al. Genes Dev.* **15,** 1140–1151 (2001).
11. You, S.-H. *et al. Nature Struct. Mol. Biol.* **20,** 182–187 (2013).
12. Gutschner, T. *et al. Cancer Res.* **73,** 1180–1189 (2013).
13. Gutschner, T. & Diederichs, S. *RNA Biol.* **9,** 703–719 (2012).

ARTICLE

doi:10.1038/nature14447

Complex archaea that bridge the gap between prokaryotes and eukaryotes

Anja Spang[1]*, Jimmy H. Saw[1]*, Steffen L. Jørgensen[2]*, Katarzyna Zaremba-Niedzwiedzka[1]*, Joran Martijn[1], Anders E. Lind[1], Roel van Eijk[1]†, Christa Schleper[2,3], Lionel Guy[1,4] & Thijs J. G. Ettema[1]

The origin of the eukaryotic cell remains one of the most contentious puzzles in modern biology. Recent studies have provided support for the emergence of the eukaryotic host cell from within the archaeal domain of life, but the identity and nature of the putative archaeal ancestor remain a subject of debate. Here we describe the discovery of 'Lokiarchaeota', a novel candidate archaeal phylum, which forms a monophyletic group with eukaryotes in phylogenomic analyses, and whose genomes encode an expanded repertoire of eukaryotic signature proteins that are suggestive of sophisticated membrane remodelling capabilities. Our results provide strong support for hypotheses in which the eukaryotic host evolved from a bona fide archaeon, and demonstrate that many components that underpin eukaryote-specific features were already present in that ancestor. This provided the host with a rich genomic 'starter-kit' to support the increase in the cellular and genomic complexity that is characteristic of eukaryotes.

Cellular life is currently classified into three domains: Bacteria, Archaea and Eukarya. Whereas the cytological properties of Bacteria and Archaea are relatively simple, eukaryotes are characterized by a high degree of cellular complexity, which is hard to reconcile given that most hypotheses assume a prokaryote-to-eukaryote transition[1,2]. In this context, it seems particularly difficult to account for the suggested presence of the endomembrane system, the nuclear pores, the spliceosome, the ubiquitin protein degradation system, the RNAi machinery, the cytoskeletal motors and the phagocytotic machinery in the last eukaryotic common ancestor (ref. 3 and references therein). Ever since the recognition of the archaeal domain of life by Carl Woese and co-workers[4,5], Archaea have featured prominently in hypotheses for the origin of eukaryotes, as eukaryotes and Archaea represented sister lineages in Woese's 'universal tree'[5]. The evolutionary link between Archaea and eukaryotes was further reinforced through studies of the transcription machinery[6] and the first archaeal genomes[7], revealing that many genes, including the core of the genetic information-processing machineries of Archaea, were more similar to those of eukaryotes[8] rather than to Bacteria. During the early stages of the genomic era, it also became apparent that eukaryotic genomes were chimaeric by nature[8,9], comprising genes of both archaeal and bacterial origin, in addition to genes specific to eukaryotes. Yet, whereas many of the bacterial genes could be traced back to the alphaproteobacterial progenitor of mitochondria, the nature of the lineage from which the eukaryotic host evolved remained obscure[1,10–13]. This lineage might either descend from a common ancestor shared with Archaea (following Woese's classical three-domains-of-life tree[5]), or have emerged from within the archaeal domain (so-called archaeal host or eocyte-like scenarios[1,14–17]). Recent phylogenetic analyses of universal protein data sets have provided increasing support for models in which eukaryotes emerge as sister to or from within the archaeal 'TACK' superphylum[18–22], a clade originally comprising the archaeal phyla Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota[23]. In support of this relationship, comparative genomics analyses have revealed several eukaryotic signature proteins (ESPs)[24] in TACK lineages, including distant archaeal homologues of actin[25] and tubulin[26], archaeal cell division proteins related to the eukaryotic endosomal sorting complexes required for transport (ESCRT)-III complex[27], and several information-processing proteins involved in transcription and translation[2,17,23]. These findings suggest an archaeal ancestor of eukaryotes that might have been more complex than the archaeal lineages identified thus far[2,23,28]. Yet, the absence of missing links in the prokaryote-to-eukaryote transition currently precludes detailed predictions about the nature and timing of events that have driven the process of eukaryogenesis[1,2,17,28]. Here we describe the discovery of a new archaeal lineage related to the TACK superphylum that represents the nearest relative of eukaryotes in phylogenomic analyses, and intriguingly, its genome encodes many eukaryote-specific features, providing a unique insight in the emergence of cellular complexity in eukaryotes.

Genomic exploration of new TACK archaea

While surveying microbial diversity in deep marine sediments influenced by hydrothermal activity from the Arctic Mid-Ocean Ridge, 16S rRNA gene sequences belonging to uncultivated archaeal candidate lineages were identified in a gravity core (GC14) sampled approximately 15 km north-northwest of the active venting site Loki's Castle[29] at 3283 m below sea level (73.763167 N, 8.464000 E) (Fig. 1a)[30,31]. Subsequent phylogenetic analyses of these sequences, which comprised ~10% of the obtained 16S reads, revealed that they belonged to the gamma clade of the Deep-Sea Archaeal Group/Marine Benthic Group B (hereafter referred to as DSAG)[31–33] (Fig. 1b–d and Supplementary Figs 1 and 2), a clade proposed to be deeply-branching in the TACK superphylum[23]. DSAG constitutes one of the most abundant and widely distributed archaeal groups in the deep marine biosphere, but so far none of its representatives have been cultured or sequenced[31].

To obtain genomic information for this archaeal lineage, we applied deep metagenomic sequencing to the GC14 sediment sample, resulting in a smaller (LCGC14, 8.6 Gbp) and a larger, multiple-strand displacement amplified (MDA) metagenome data set (LCGC14AMP, 56.6 Gbp; Fig. 2a; Supplementary Fig. 3 and Supplementary Table 1). Given the

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden. [2]Department of Biology, Centre for Geobiology, University of Bergen, N-5020 Bergen, Norway. [3]Division of Archaea Biology and Ecogenomics, Department of Ecogenomics and Systems Biology, University of Vienna, A-1090 Vienna, Austria. [4]Department of Medical Biochemistry and Microbiology, Uppsala University, SE-75123 Uppsala, Sweden. †Present address: Groningen Institute for Evolutionary Life Sciences, University of Groningen, NL-9747AG Groningen, The Netherlands. *These authors contributed equally to this work.

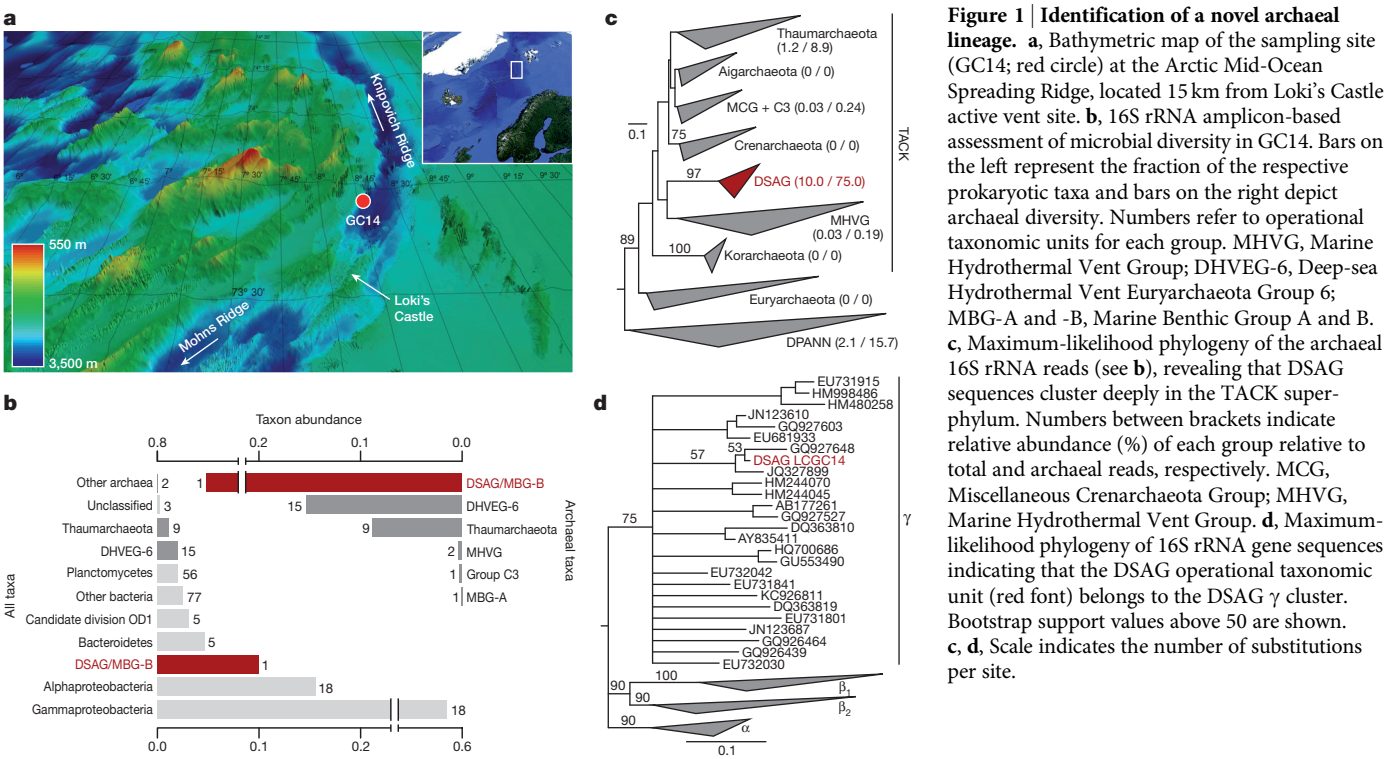


Figure 1 | Identification of a novel archaeal lineage. **a**, Bathymetric map of the sampling site (GC14; red circle) at the Arctic Mid-Ocean Spreading Ridge, located 15 km from Loki's Castle active vent site. **b**, 16S rRNA amplicon-based assessment of microbial diversity in GC14. Bars on the left represent the fraction of the respective prokaryotic taxa and bars on the right depict archaeal diversity. Numbers refer to operational taxonomic units for each group. MHVG, Marine Hydrothermal Vent Group; DHVEG-6, Deep-sea Hydrothermal Vent Euryarchaeota Group 6; MBG-A and -B, Marine Benthic Group A and B. **c**, Maximum-likelihood phylogeny of the archaeal 16S rRNA reads (see **b**), revealing that DSAG sequences cluster deeply in the TACK superphylum. Numbers between brackets indicate relative abundance (%) of each group relative to total and archaeal reads, respectively. MCG, Miscellaneous Crenarchaeota Group; MHVG, Marine Hydrothermal Vent Group. **d**, Maximum-likelihood phylogeny of 16S rRNA gene sequences indicating that the DSAG operational taxonomic unit (red font) belongs to the DSAG γ cluster. Bootstrap support values above 50 are shown. **c**, **d**, Scale indicates the number of substitutions per site.

deeper coverage, the latter data set was used to extract marker genes that carry an evolutionary coherent phylogenetic signal (Supplementary Tables 2 and 3). Using single gene phylogenies of these markers, contigs attributable to either one of the archaeal lineages present in the LCGC14AMP metagenome (DSAG, DSAG-related, DPANN and Thaumarchaeota), could be extracted. These taxon-specific contigs were used as training sets for supervised binning of contigs present in both the LCGC14 and LCGC14AMP metagenomes (Supplementary Fig. 4). This approach resulted in the identification of two DSAG bins (from LCGC14 and LCGC14AMP, respectively) as well as one DSAG-related bin (bin Loki2/3 from LCGC14AMP). We focused on the DSAG bin from the non-amplified data set to avoid potential biases introduced by MDA (see Methods). The analyses of the low-abundant DSAG-related lineages were based on the MDA-amplified LCGC14AMP data set.

After removal of small (<1 kbp) and low-coverage contigs (Supplementary Fig. 5), reads mapping to the remaining DSAG bin contigs were reassembled into 504 contigs, yielding a 92% complete, 1.4 fold-redundant composite genome ('Lokiarchaeum') of 5.1 Mbp, which encodes 5,381 protein coding genes as well as single copies of the 16S and 23S rRNA genes (Supplementary Table 4 and Supplementary Discussion 1). The DSAG-related bin (Loki2/3 from LCGC14AMP) was found to contain two low-abundant, distinct lineages, displaying slight but marked differences in GC content of 32.8 and 29.9%, allowing for separation into two distinct groups (Loki2 and Loki3) (Supplementary Fig. 6). Since these two lineages represent low-abundance community members, only partial genomes could be recovered. The Loki2/3 contigs did not contain 16S rRNA genes, rendering it impossible to attribute them to any of the uncultured archaeal 16S phylotypes identified in the GC14 sediments, such as the low-abundance Marine Hydrothermal Vent Group archaea (abundance ~0.05%; Fig. 1c). However, phylogenetic marker genes were extracted for these lineages as well (21 and 34 markers for Loki2 and Loki3, respectively) since their inclusion was potentially useful in resolving the phylogenetic placement of the Lokiarchaeum lineage.

Lokiarchaeota and Eukarya are monophyletic

To determine the phylogenetic affiliation of Lokiarchaeum and the Loki2/Loki3 lineages, maximum-likelihood and Bayesian inference phylogenetic analyses were performed, using sophisticated models of molecular sequence evolution. By implementing relaxed assumptions of homogeneous amino acid composition across sites or across branches of the tree, these models are less sensitive to long-branch attraction and other phylogenetic artefacts. Both maximum-likelihood and Bayesian inference analyses of concatenated alignments comprising 36 conserved phylogenetic marker proteins[20] (Supplementary Tables 2 and 3) revealed that the DSAG and DSAG-related archaea (hereafter referred to as 'Lokiarchaeota') represent a monophyletic, deeply branching clade of the TACK superphylum. Loki3 represented the deepest branch of the Lokiarchaeota, and Lokiarchaeum and Loki2 were inferred to be sister lineages with maximum support (Supplementary Fig. 7). Intriguingly, when eukaryotes were included in our phylogenetic analyses, they were confidently positioned within the Lokiarchaeota (posterior probability = 1; bootstrap support = 80; Fig. 2b; Supplementary Figs 8 and 9), as the sister group of the Loki3 lineage (Fig. 2b). Robust assessment of these phylogenetic inferences (Supplementary Figs 10–14 and Supplementary Table 5) revealed strong support for the Lokiarchaeota–Eukarya affiliation (Supplementary Discussion 2).

The proposed naming of the Eukarya-affiliated candidate phylum Lokiarchaeota and the Lokiarchaeum lineage is made in reference to the sampling location, Loki's Castle[29], which in turn was named after the Norse mythology's shape-shifting deity Loki. Loki has been described as "a staggeringly complex, confusing, and ambivalent figure who has been the catalyst of countless unresolved scholarly controversies"[34], in analogy to the ongoing debates on the origin of eukaryotes.

Presence of diverse and abundant ESPs

As our phylogenetic analyses strongly support a common ancestry of Lokiarchaeota and eukaryotes, we investigated the presence of putative ESPs[24] in the composite Lokiarchaeum genome. The amount of genomic data obtained for the Loki2/3 lineages was too low to perform detailed gene content analyses. A comparative taxonomic assessment of the Lokiarchaeum composite proteome revealed that a large fraction (32%) of its proteins displayed no significant similarity to any known protein, and that roughly as many proteins display highest similarity to archaeal and bacterial proteins (26% and 29%,

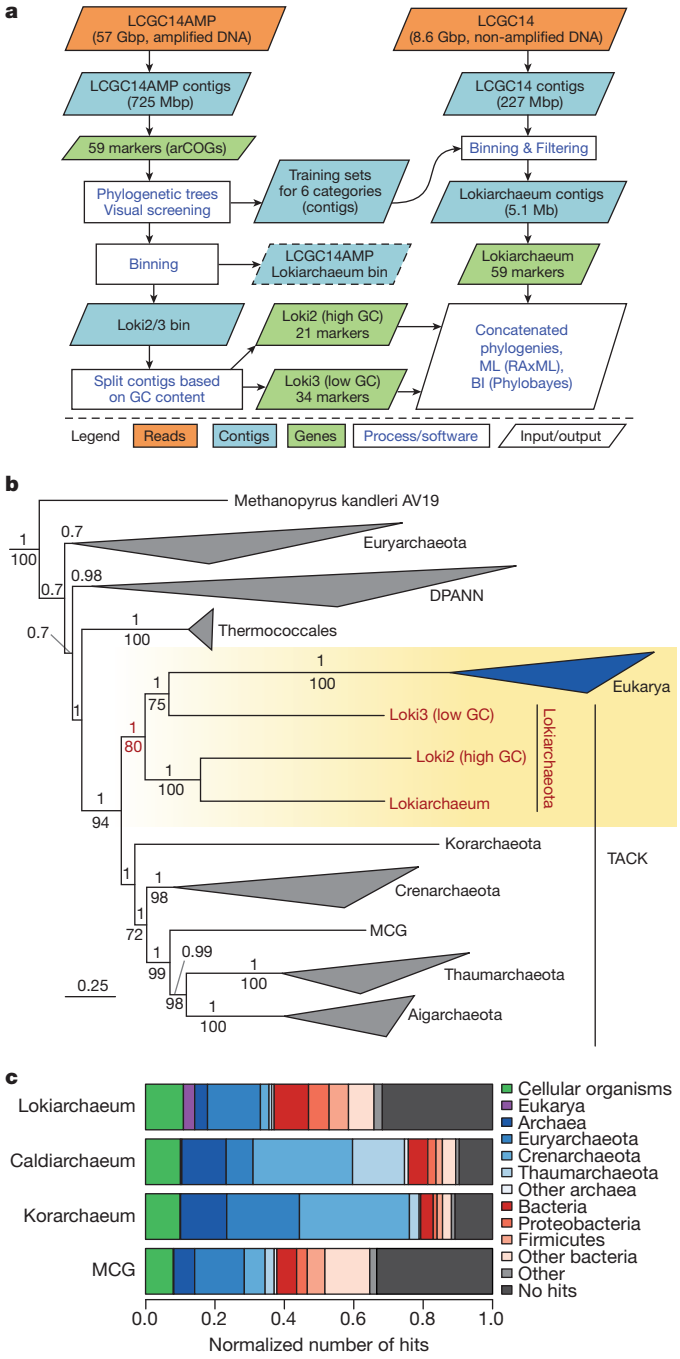


Figure 2 | Metagenomic reconstruction and phylogenetic analysis of Lokiarchaeum. **a**, Schematic overview of the metagenomics approach. BI, Bayesian inference; ML, maximum likelihood. **b**, Bayesian phylogeny of concatenated alignments comprising 36 conserved phylogenetic marker proteins using sophisticated models of protein evolution (Methods), showing eukaryotes branching within Lokiarchaeota. Numbers above and below branches refer to Bayesian posterior probability and maximum-likelihood bootstrap support values, respectively. Posterior probability values above 0.7 and bootstrap support values above 70 are shown. Scale indicates the number of substitutions per site. **c**, Phylogenetic breakdown of the Lokiarchaeum proteome, in comparison with proteomes of Korarchaeota, Aigarchaeota (Caldiarchaeum) and Miscellaneous Crenarchaeota Group (MCG) archaea. Category 'Other' contains proteins assigned to the root of cellular organisms, to viruses and to unclassified proteins.

respectively; Fig. 2c and Supplementary Fig. 15), which is in accordance with recent findings that suggest major inter-domain gene exchange between Bacteria and Archaea[35,36] (Supplementary Discussion 3). Most notably, a significant part of the predicted proteome (175 proteins or 3.3%) was most similar to eukaryotic proteins (Fig. 2c) and revealed a dominance of proteins, which in eukaryotes are involved in membrane deformation and cell shape formation processes, including phagocytosis[37] (Extended Data Table 1 and Supplementary Table 6). Several lines of evidence support that the presence of these proteins is not the result of potential contaminating eukaryotic sequence data. First, genes encoding Lokiarchaeum ESPs and other proteins most similar to eukaryotes were always flanked by prokaryotic genes (Supplementary Fig. 16), and most were encoded by contigs that also contained archaeal signature genes. Second, ESP-encoding contigs displayed high (>20×) read coverage, while eukaryotic sequences could not be detected in the LCGC14 data set, and represented only a negligible fraction of the LCGC14AMP metagenome. Furthermore, the amplicon data generated with universal 16S/18S primers did not reveal any 18S rRNA genes of eukaryotic origin (Fig. 1b). Third, phylogenetic analyses of several Lokiarchaeal ESPs revealed their emergence at the base of eukaryotic clades (see below), indicating that these proteins represent archaeal out-groups of the eukaryotic proteins rather than being truly eukaryotic in origin. Fourth, Lokiarchaeum appears to contain bona fide archaeal informational processing machineries (Supplementary Discussion 4 and Supplementary Tables 7–9) and, irrespective of the significant amount of ESPs in its genome, lacks many other key eukaryotic features. Finally, we could also identify highly similar homologues of the Lokiarchaeal ESPs in a recent and independently generated marine sediment metagenome derived from a sediment core sample off the Shimokita Peninsula of Japan, in which DSAG comprises a significant part of the microbial community[38]. As the function and evolution of the Lokiarchaeal ESPs hold relevance for understanding the origin of the eukaryotic cell, we review some of the key findings in more detail below.

Potential dynamic actin cytoskeleton

Actins represent key structural proteins of eukaryotic cells and comprise filaments that are crucial for various cellular processes, including cell division, motility, vesicle trafficking and phagocytosis[39]. The Lokiarchaeum genome encodes five actin homologues that display higher similarity to eukaryotic actins and actin-related proteins (ARPs) than to crenactins, a group of archaeal actin homologues that were recently shown to be involved in cell shape formation[25,37,40] (Supplementary Table 6). This observation was confirmed in a phylogenetic analysis of the Lokiarchaeal actins that also included homologues identified in a recently published marine sediment metagenome[38] (up to 99% identity) as well as in the LCGC14 and LCGC14AMP metagenomes (Fig. 3a and Supplementary Fig. 17). Lokiarchaeal actins ('Lokiactins') comprise several distinct clusters, some of which branch at the base of distinctive eukaryotic actin and ARP clusters, albeit with weak support (Fig. 3a). Despite the poor resolution of several deeper nodes in the actin tree, strong support is provided for a common ancestry of Lokiactins and eukaryotic actins, indicating that the proliferation of actins already occurred in the archaeal ancestor of eukaryotes. Notably, the Lokiarchaeum genome also encodes several hypothetical short proteins containing gelsolin-like domains that so far appear to be absent from bacterial and any other archaeal genomes (Extended Data Table 1, Supplementary Tables 6 and 10 and Supplementary Discussion 5). In eukaryotes, these protein domains are part of the villin/gelsolin superfamily of proteins, which comprise various key regulators of actin filament assembly and disassembly[41]. Although the function of these hypothetical gelsolin-domain proteins remains to be elucidated, it is tempting to speculate that Lokiarchaeum has a dynamic actin cytoskeleton.

Genomic expansion of small GTPases

Small GTPases belonging to the Ras superfamily comprise one of the largest protein families in eukaryotes, where they are involved in various regulatory processes, including cytoskeleton remodelling, signal transduction, nucleocytoplasmic transport and vesicular trafficking[42]. Being

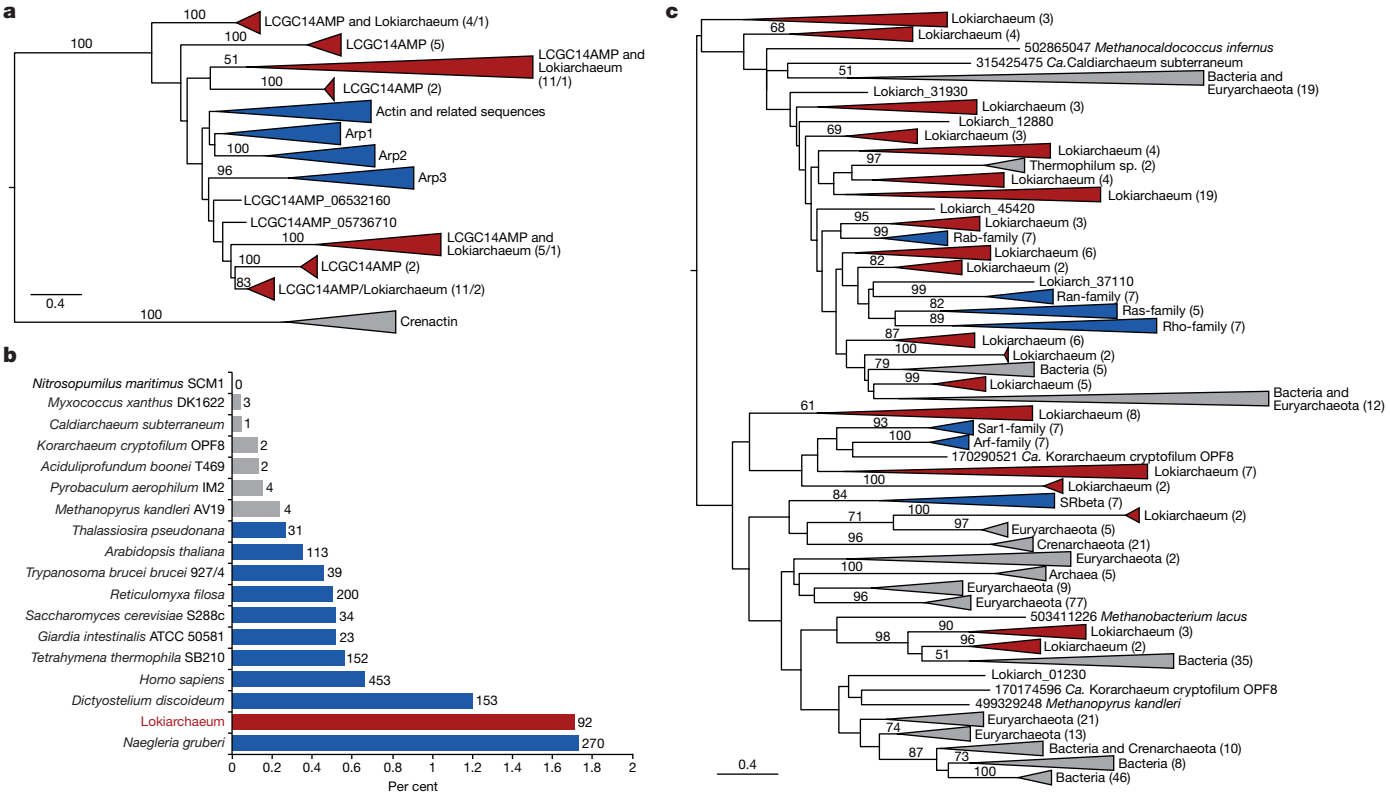


Figure 3 | Identification and phylogeny of small GTPases and actin orthologues. **a**, Maximum-likelihood phylogeny of 378 aligned amino acid residues of actin homologues identified in Lokiarchaeum and in the LCGC14AMP metagenome, including eukaryotic actins, ARP1–3 homologues and crenactins[25]. Consecutive numbers in brackets refer to the number of sequences in a respective clade from LCGC14AMP and Lokiarchaeum, respectively. **b**, Relative amount of small GTPases (assigned to IPR006689 and IPR001806) in the Lokiarchaeum genome in comparison with other eukaryotic, archaeal and bacterial species. Numbers refer to total amount of small GTPases per predicted proteome. **c**, Maximum-likelihood phylogeny of 150 aligned amino acid residues of small Ras- and Arf-type GTPases (IPR006689 and IPR001806) in all domains of life. Numbers in brackets refer to the number of sequences in the respective clades. **a**, **c**, Sequence clusters comprising Lokiarchaeum and/or LCGC14AMP sequences (red), eukaryotes (blue) and Bacteria/Archaea (grey) have been collapsed. Bootstrap values above 50 are shown. Scale indicates the number of substitutions per site.

key regulators of actin cytoskeleton dynamics, these small GTPases represent essential components for the process of phagocytosis in eukaryotes. Intriguingly, the analysis of Lokiarchaeal ESPs revealed a multitude of Ras-superfamily GTPases, comprising nearly 2% of the Lokiarchaeal proteome (Fig. 3b). The relative amount of small GTPases in the Lokiarchaeum genome is comparable to that observed in several unicellular eukaryotes, only being surpassed by the protist *Naegleria gruberi*. In contrast, bacterial and archaeal genomes encode only few, if any, small GTPase homologues of the Ras superfamily (Fig. 3b).

Phylogenetic analyses of the Lokiarchaeal small GTPases revealed that these represent several distinct clusters, each of which comprises several GTPase sequences (Fig. 3c and Supplementary Fig. 18). Although phylogenetic analyses failed to resolve most of the deeper nodes, several of the eukaryotic small GTPase families appear to share a common ancestry with Lokiarchaeal GTPases (Fig. 3c), suggesting an archaeal origin of specific subgroups of the eukaryotic small GTPases, followed by independent expansions in eukaryotes and Lokiarchaeota. This scenario contrasts with previous studies that have suggested that eukaryotic small GTPases were acquired from the alphaproteobacterial progenitor of mitochondria[37].

Although genes encoding canonical eukaryotic GTPase-activating proteins (GAPs) were absent in Lokiarchaeota, twelve roadblock/LC7-domain-containing proteins were identified (Supplementary Tables 6 and 10). While such proteins have been implicated in dynein organization in eukaryotes, roadblock/LC7 protein MglB of the bacterium *Myxococcus xanthus* was shown to act as a GAP of the small GTPase MglA[43]. Hence, the Lokiarchaeal roadblock/LC7 proteins represent possible candidates for alternative GAPs in this archaeon.

Presence of a primordial ESCRT complex

In eukaryotes, the ESCRT machinery represents an essential component of the multivesicular endosome pathway for lysosomal degradation of damaged or superfluous proteins, and it plays a role in several budding processes including cytokinesis, autophagy and viral budding[44]. The ESCRT machinery generally consists of the ESCRT-I–III subcomplexes, as well as associated subunits[45]. The analysis of the Lokiarchaeum genome revealed the presence of an ESCRT gene cluster (Fig. 4a), as well as of several additional proteins homologous to components of the eukaryotic multivesicular endosome pathway. For instance, Lokiarchaeum encodes divergent SNF7 domain proteins of the eukaryotic ESCRT-III complex, which appear to represent members of the Vps2/Vps24/Vps46 and Vps20/Vps32/Vps60 families, respectively. A phylogenetic analysis of the Lokiarchaeal SNF7 domain proteins revealed that these branch at the base of these two eukaryotic ESCRT-III families with low bootstrap support (Fig. 4b and Supplementary Fig. 19), not only indicating that they might represent ancestral SNF7 copies, but also suggesting that the last eukaryotic common ancestor already inherited two divergent SNF7-domain-encoding genes from its putative archaeal ancestor rather than a single gene[46]. Furthermore, the gene cluster encodes an ATPase that displays closest resemblance to eukaryotic VPS4-type ATPases, including katanin, membrane scaffold protein (MSP) and spastin (Fig. 4c and Supplementary Fig. 20) as well as hypothetical proteins that show significant similarity

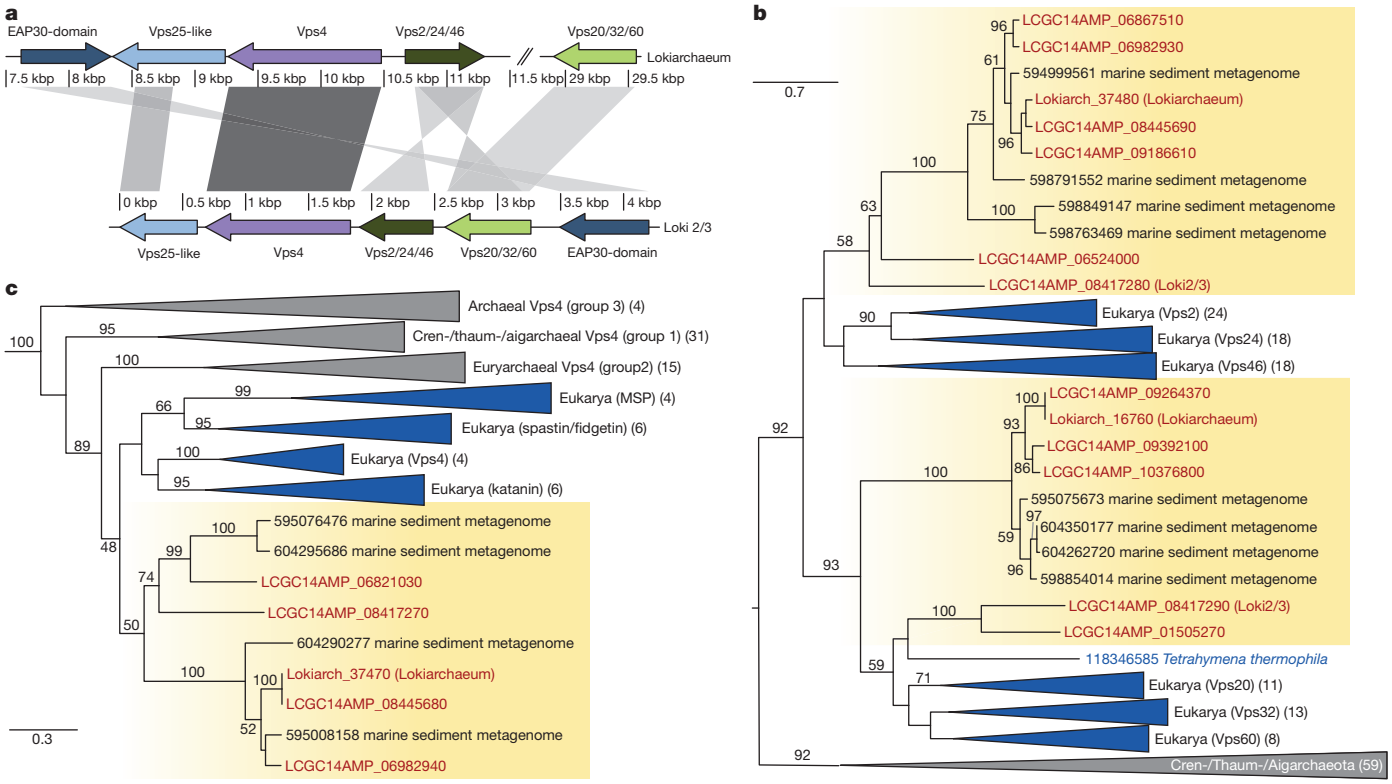


Figure 4 | Identification of ESCRT components in the Lokiarchaeum genome. **a**, Schematic overview of ESCRT gene clusters identified in Lokiarchaeum and Loki2/3. Intensity of shading between homologous sequences is correlated with BLAST bit score. **b**, Maximum-likelihood phylogeny of 207 aligned amino acid residues of ESCRT-III homologues identified in Lokiarchaeum, LCGC14AMP and other archaeal lineages. Eukaryotic homologues include the two distantly related families Vps2/24/46 and Vps20/32/60. Bootstrap support values above 50 are shown. **c**, Maximum-likelihood phylogeny of 388 aligned amino acid residues of AAA-type Vps4 ATPases including representatives for each of the four major eukaryotic sub-groups (membrane scaffold protein (MSP), katanin, spastin/fidgetin and Vps4) as well as homologues identified in the Lokiarchaeum genome, in LCGC14AMP and in sequenced archaeal genomes. Bootstrap support values below 45 are not shown. **b**, **c**, Scale indicates the number of substitutions per site. Numbers in brackets refer to the number of sequences in the respective clades.

to EAP30-domain-containing proteins (Vps36/22) and Vps25, respectively (Fig. 4a; Supplementary Figs 21 and 22). In eukaryotes, Vps22, Vps25 and Vps36 are components of the ESCRT-II complex, which comprises two to three of these proteins depending on the eukaryotic species[46]. In addition, a protein domain analysis of the Lokiarchaeum proteome identified a Vps28-like protein, a component of the eukaryotic ESCRT-I subcomplex that links the ubiquitin pathway to vesicular transport and which, apart from Vps28, comprises Vps23 and Vps37 (Extended Data Table 1 and Supplementary Fig. 23). The different subunits of the eukaryotic ESCRT-I complex share similar two-helix core domains and have been suggested to have evolved from a single ancestral sequence[47], which we now propose to be of archaeal origin.

Finally, the Lokiarchaeum proteome was found to contain hypothetical proteins containing Longin-like domains, as well as several proteins belonging to the BAR/IMD superfamily (Supplementary Tables 6 and 10), comprising curvature sensing protein families involved in various aspects of vesicle/membrane trafficking or remodelling processes in eukaryotes. These findings suggest that Lokiarchaeum contains a primordial version of a eukaryotic ESCRT vesicle trafficking complex. In eukaryotes, ubiquitylation of target proteins represents a critical step in ESCRT-mediated protein degradation through the multivesicular endosome pathway[44,48]. The Lokiarchaeum genome contains a gene cluster that encodes several components required for a functional ubiquitin modifier system, including homologues for ubiquitin-activating enzyme E1, ubiquitin-conjugating enzyme E2, and 26S proteasome regulatory subunit RPN11. In addition, several hypothetical proteins with ubiquitin-like domains were identified in Lokiarchaeum, as well as diverse zinc-finger/RING-domain-containing proteins, some of which might serve as candidates for E3 ubiquitin protein ligases (Supplementary Tables 6 and 10). Several of these components have also been identified in Aigarchaeota[49].

A 'complex' archaeal ancestor of Eukarya

We have identified and characterized the genome of Lokiarchaeota, a novel, deeply rooting clade of the archaeal TACK superphylum, which in phylogenomic analyses of universal proteins forms a monophyletic group with eukaryotes. While the obtained phylogenomic resolution testifies to a deep archaeal ancestry of eukaryotes, the Lokiarchaeum genome content holds valuable clues about the nature of the archaeal ancestor of eukaryotes, and about the process of eukaryogenesis. Many of the ESPs previously identified in different TACK lineages are united in Lokiarchaeum, indicating that the patchy distribution of ESPs amongst archaea is most likely the result of lineage-specific losses[2] (Fig. 5). Moreover, the Lokiarchaeum genome significantly expands the total number of ESPs in Archaea, lending support to the observed phylogenetic affiliation of Lokiarchaeota and eukaryotes. Finally, and importantly, sequence-based functional predictions for these new ESPs indicate a predominance of proteins that play pivotal roles in various membrane remodelling and vesicular trafficking processes in eukaryotes. It is also noteworthy that Lokiarchaeum appears to encode the most 'eukaryotic-like' ribosome identified in Archaea thus far (Supplementary Discussion 4), including a putative homologue of eukaryotic ribosomal protein L22e (Fig. 5; Supplementary Fig. 24 and Supplementary Tables 7 and 8).

Taken together, our data indicate that the archaeal ancestor of eukaryotes was even more complex than previously inferred[2] and allow us to

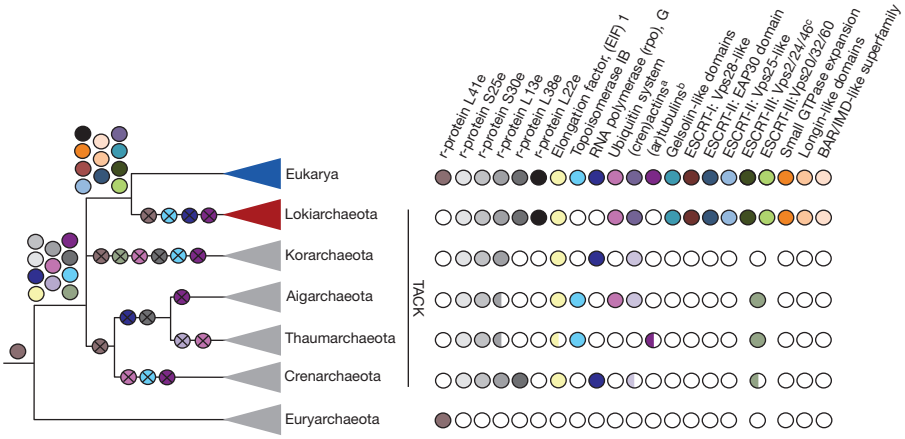


Figure 5 | The complex archaeal ancestry of eukaryotes. Schematic overview of the distribution of ESPs in major archaeal lineages across the tree of life. Each ESP is depicted as a coloured circle and losses are indicated with a cross. Patchy distribution and absence of a particular ESP in archaeal phyla is indicated by half-shaded and white circles, respectively. [a]While eukaryotes and Lokiarchaeota contain bona fide actins, other archaea encode the more distantly related Crenactins. [b]Only few members of the Thaumarchaeota contain distantly related homologs of tubulins (ar-tubulins). [c]Thaum-, Aig- and some Crenarchaeota contain distant homologues of ESCRT-III (SNF7 domain proteins).

speculate on the timing and order of several key events in the process of eukaryogenesis. For example, the identification of archaeal genes involved in membrane remodelling and vesicular trafficking processes indicates that the emergence of cellular complexity was already underway before the acquisition of the mitochondrial endosymbiont, which now appears to be a universal feature of all eukaryotes[28,37,50]. Indeed, based upon our results it seems plausible that the archaeal ancestor of eukaryotes had a dynamic actin cytoskeleton and potentially endo- and/or phagocytic capabilities, which would have facilitated the invagination of the mitochondrial progenitor.

The present identification and genomic characterization of a novel archaeal group that shares a common ancestry with eukaryotes indicates that the gap between prokaryotes and eukaryotes might, to some extent, be a result of poor sampling of the existing archaeal diversity. Environmental surveys have revealed the existence of a plethora of uncultured archaeal lineages, and some of these likely represent even closer relatives of eukaryotes. Excitingly, the genomic exploration of these archaeal lineages has now come within reach. Such endeavours, combined with prospective studies focusing on uncovering metabolic, chemical and cell biological properties of these lineages, will uncover further details about the identity and nature of the archaeal ancestor of eukaryotes, shedding new light on the evolutionary dark ages of the eukaryotic cell.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 December 2014; accepted 1 April 2015.
Published online 6 May 2015.

1. Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440,** 623–630 (2006).
2. Koonin, E. V. & Yutin, N. The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6,** a016188 (2014).
3. Koumandou, V. L. *et al.* Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48,** 373–396 (2013).
4. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74,** 5088–5090 (1977).
5. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87,** 4576–4579 (1990).
6. Pühler, G. *et al.* Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl Acad. Sci. USA* **86,** 4569–4573 (1989).
7. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273,** 1058–1073 (1996).
8. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* **95,** 6239–6244 (1998).
9. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nature Rev. Microbiol.* **12,** 449–455 (2014).
10. Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P. & Brochier-Armanet, C. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Rev. Microbiol.* **8,** 743–752 (2010).
11. Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. & Koonin, E. V. The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25,** 1619–1630 (2008).
12. Rochette, N. C., Brochier-Armanet, C. & Gouy, M. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Mol. Biol. Evol.* **31,** 832–845 (2014).
13. Thiergart, T., Landan, G., Schenk, M., Dagan, T. & Martin, W. F. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol. Evol.* **4,** 466–485 (2012).
14. Henderson, E. *et al.* A new ribosome structure. *Science* **225,** 510–512 (1984).
15. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11,** 209 (2010).
16. Lake, J. A. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331,** 184–186 (1988).
17. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
18. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaebacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105,** 20356–20361 (2008).
19. Foster, P. G., Cox, C. J. & Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Phil. Trans. R. Soc. Lond. B* **364,** 2197–2207 (2009).
20. Guy, L., Saw, J. H. & Ettema, T. J. The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6,** a016022 (2014).
21. Lasek-Nesselquist, E. & Gogarten, J. P. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* **69,** 17–38 (2013).
22. Williams, T. A., Foster, P. G., Nye, T. M., Cox, C. J. & Embley, T. M. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. Lond. B* **279,** 4870–4879 (2012).
23. Guy, L. & Ettema, T. J. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19,** 580–587 (2011).
24. Hartman, H. & Fedorov, A. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl Acad. Sci. USA* **99,** 1420–1425 (2002).
25. Ettema, T. J., Lindås, A.-C. & Bernander, R. An actin-based cytoskeleton in archaea. *Mol. Microbiol.* **80,** 1052–1061 (2011).
26. Yutin, N. & Koonin, E. V. Archaeal origin of tubulin. *Biol. Direct* **7,** 10 (2012).
27. Lindås, A.-C., Karlsson, E. A., Lindgren, M. T., Ettema, T. J. & Bernander, R. A unique cell division machinery in the Archaea. *Proc. Natl Acad. Sci. USA* **105,** 18942–18946 (2008).
28. Martijn, J. & Ettema, T. J. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem. Soc. Trans.* **41,** 451–457 (2013).
29. Pedersen, R. B. *et al.* Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat. Commun.* **1,** 126 (2010).
30. Jørgensen, S. L. *et al.* Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl Acad. Sci. USA* **109,** E2846–E2855 (2012).
31. Jørgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the Arctic mid-ocean spreading ridge. *Front. Microbiol.* **4,** 299 (2013).
32. Inagaki, F. *et al.* Microbial communities associated with geological horizons in coastal subseafloor sediments from the Sea of Okhotsk. *Appl. Environ. Microbiol.* **69,** 7224–7235 (2003).
33. Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A. & Reysenbach, A. L. Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl. Environ. Microbiol.* **65,** 4375–4384 (1999).
34. Von Schnurbein, S. The function of Loki in Snorri Sturluson's *Edda*. *Hist. Relig.* **40,** 109–124 (2000).
35. Deschamps, P., Zivanovic, Y., Moreira, D., Rodriguez-Valera, F. & Lopez-Garcia, P. Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome Biol. Evol.* **6,** 1549–1563 (2014).
36. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517,** 77–80 (2015).
37. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4,** 9 (2009).
38. Kawai, M. *et al.* High frequency of phylogenetically diverse reductive dehalogenase-homologous genes in deep subseafloor sedimentary metagenomes. *Front. Microbiol.* **5,** 80 (2014).

39. Pollard, T. D. & Cooper, J. A. Actin, a central player in cell shape and movement. *Science* **326,** 1208–1212 (2009).
40. Bernander, R., Lind, A. E. & Ettema, T. J. An archaeal origin for the actin cytoskeleton: Implications for eukaryogenesis. *Commun. Integr. Biol.* **4,** 664–667 (2011).
41. Pollard, T. D. & Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112,** 453–465 (2003).
42. Takai, Y., Sasaki, T. & Matozaki, T. Small GTP-binding proteins. *Physiol. Rev.* **81,** 153–208 (2001).
43. Zhang, Y., Franco, M., Ducret, A. & Mignot, T. A bacterial Ras-like small GTP-binding protein and its cognate GAP establish a dynamic spatial polarity axis to control directed motility. *PLoS Biol.* **8,** e1000430 (2010).
44. Hurley, J. H. The ESCRT complexes. *Crit. Rev. Biochem. Mol. Biol.* **45,** 463–487 (2010).
45. Field, M. C. & Dacks, J. B. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Curr. Opin. Cell Biol.* **21,** 4–13 (2009).
46. Leung, K. F., Dacks, J. B. & Field, M. C. Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage. *Traffic* **9,** 1698–1716 (2008).
47. Kostelansky, M. S. *et al.* Structural and functional organization of the ESCRT-I trafficking complex. *Cell* **125,** 113–126 (2006).
48. Raiborg, C. & Stenmark, H. The ESCRT machinery in endosomal sorting of ubiquitylated membrane proteins. *Nature* **458,** 445–452 (2009).
49. Nunoura, T. *et al.* Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res.* **39,** 3204–3223 (2011).
50. Poole, A. M. & Gribaldo, S. Eukaryotic origins: how and when was the mitochondrion acquired? *Cold Spring Harb. Perspect. Biol.* **6,** a015990 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We dedicate this paper to the memory of Rolf Bernander. We thank P. Offre and I. de Bruijn for technical advice and useful discussions, and A. Denny for image processing. We also acknowledge the help from chief scientist R. B. Pedersen, the scientific party and the entire crew on board the Norwegian research vessel G.O. Sars during the summer 2010 expedition. All sequencing was performed by the National Genomics Infrastructure sequencing platforms at the Science for Life Laboratory at Uppsala University, a national infrastructure supported by the Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. We thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at Uppsala University and the Swedish National Infrastructure for Computing (SNIC) at the PDC Center for High-Performance Computing for providing computational resources. This work was supported by grants of the Swedish Research Council (VR grant 621-2009-4813), the European Research Council (ERC Starting grant 310039-PUZZLE_CELL) and the Swedish Foundation for Strategic Research (SSF-FFL5) to T.J.G.E., and by grants of the Carl Tryggers Stiftelse för Vetenskaplig Forskning (to A.S.), the Wenner-Gren Stiftelserna (to J.H.S.), and by Marie Curie IIF (331291 to J.H.S.) and IEF (625521 to A.S.) grants by the European Union. S.L.J received financial support from the H2DEEP project through the EuroMARC program, and by the Research Council of Norway through the Centre for Geobiology, University of Bergen. C.S. is supported by the Austrian Science Fund (FWF grant P27017).

Author Contributions T.J.G.E., S.L.J. and C.S. conceived the study. S.L.J. provided deep-sea sediments and isolated community DNA. R.v.E., J.H.S. and A.E.L. prepared sequencing libraries. A.E.L., J.H.S., S.L.J. and J.M. analysed environmental sequence data. L.G., K.Z.-N. and J.H.S. performed, optimised and analysed metagenomic sequence assemblies. L.G., J.H.S., A.S., K.Z.-N. and T.J.G.E. analysed genomic data and performed phylogenetic analyses. A.S., L.G., S.L.J. and T.J.G.E analysed genomic signatures of DSAG. T.J.G.E., A.S., S.L.J. and L.G. wrote, and all authors edited and approved the manuscript.

Author Information Sequence data have been deposited to the NCBI Sequence Read Archive under study number SRP045692, which includes 16 rRNA reads (experiment number SRX872366). Protein sequences of Loki2/3 were deposited to GenBank under accession numbers KP869578–KP869724. The Lokiarchaeum genome bin and the LCGC14 metagenome projects have been deposited at DDBJ/EMBL/GenBank under the accessions JYIM00000000 and LAZR00000000, respectively. The versions described in this paper are versions JYIM01000000 and LAZR01000000. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.J.G.E. (thijs.ettema@icm.uu.se) or L.G. (lionel.guy@imbim.uu.se).

METHODS

No statistical methods were used to predetermine sample size.

Sampling site and sample description. A 2-m long gravity core (GC14) was retrieved from the Arctic Mid-Ocean Ridge during summer 2010 (approximately 15 km north-northwest of the active venting site Loki's Castle; 3283 m below sea level; 73.763167 N, 8.464000 E) (Fig. 1a). Samples for geochemistry and microbiology were collected immediately and either processed on board or frozen for later analysis. Upon port arrival, the core was stored in sealed core liners at 4 °C (core depository facility, University of Bergen, Norway). Comprehensive geochemical and microbial characteristics from this and adjacent sites have been described elsewhere[51,52]. The core consists of hemipelagic-glaciomarine sediments receiving episodic hydrothermal input. The oxygen penetration depth was estimated to ~50 cm below sea floor (b.s.f.) and the content of organic carbon varied between 0.6 - 1.3%. While no measurable amounts of methane or sulphide could be measured, high and fluctuating levels of dissolved iron were detected. The relative abundance of bacterial and archaeal 16S rRNA gene copy numbers was estimated by quantitative PCR (qPCR) previously[52], indicating high abundance of the DSAG in several of the investigated sediment horizons, especially at 75 cm b.s.f. (up to 40% of the total prokaryotic population; 2.7×10^6 copies per gram sediment). Thus, sample material from horizon at 75 cm b.s.f. was used for all downstream analyses including amplicon and metagenome libraries.

DNA extraction and genomic DNA amplification. To obtain sufficient amounts of genomic DNA for sequencing library preparation, new sample material was obtained from the 75-cm-b.s.f. layer of gravity core GC14 in summer 2013. After qPCR-based verification of high DSAG abundance in the re-sampled material, DNA was extracted from 7.5 g sediment using the FastDNA spin kit for soil in conjunction with the FastPrep-24 instrument (MP Biomedicals) following manufacturer's protocol, except for the addition of polyadenosine as described in ref. 53. The individual extractions were then pooled and concentrated to a final volume of 50 µl using Amicon Ultra-0.5 filters (50.000 NMWL) following the manufacturer's protocol (Merck Millipore). Due to low yield and presence of inhibitors, 2.73 ng of this genomic DNA was amplified using the REPLI-g ultrafast mini kit (Qiagen) according to the standard protocol for purified genomic DNA.

Amplicon sequencing and analysis of 16S rDNA phylogenetic analyses. To get a better estimate of the microbial diversity of Loki's Castle sediment core LCGC14, 'universal' primer pairs (A519F (5′-CAGCMGCCGCGGTAA-3′) and U1391R (5′-ACGGGCGGTGWGTRC-3′)) were used to amplify a ~900 bp fragment of the 16S rRNA genes present in the non-amplified genomic community DNA (extracted from LCGC14, 75 cm b.s.f.) using the following conditions: 15 min of heat activation of polymerase at 95 °C and 35 cycles of 95 °C (30 s), 54 °C (45 s), 72 °C (60 s), followed by final extension at 72 °C for 7 min. Qiagen HotStar Taq DNA polymerase was used for the PCR reactions. Subsequently, PCR products of the correct size were purified with Qiagen PCR purification kit, and quantified using a Nanodrop ND-3300 fluorospectrometer (Thermo Scientific). Clean PCR products were then used as input materials for library construction using TruSeq DNA LT Sample Prep Kit (Illumina) according to the manufacturer's instructions and applied to sequencing with an Illumina MiSeq instrument. The Illumina MiSeq run produced two 300-bp paired-end reads. Raw MiSeq fastq sequences were treated with Trimmomatic tool (v0.32)[54] using the following options: TRAILING:20, MINLEN:235 and CROP:235, to remove trailing sequences below a phred quality score of 20 and to achieve uniform sequence lengths for downstream clustering processes. Remaining traces of Illumina adaptor sequences were removed by SeqPrep (https://github.com/jstjohn/SeqPrep) and by BLAST[55] searches against NCBI Univec database. Quality-filtered MiSeq reads were checked for correct orientation of the 16S rRNA sequence in the paired-end reads and those containing the forward primer sequence (A519F) were extracted for OTU clustering with UPARSE pipeline[56], setting a OTU cutoff threshold to 97%. Chimeric sequences were filtered out by the Uchime tool[57] integrated in the UPARSE pipeline. Remaining chimeric sequences, if still present, were manually checked and removed. Abundances of each OTU were calculated by mapping the chimaera-filtered OTUs against the quality-filtered reads using the UPARSE pipeline. Using the mothur package (v1.33.2)[58], representative sequences for each OTU were aligned together with the Silva NR99 release-115[59] alignment file to classify the OTUs.

Phylogenetic analysis of archaeal 16S rRNA gene sequences. Twenty-nine archaeal OTUs identified from the amplicon data were aligned together with 220 sequences representing the major clades in the archaeal 16S rRNA tree according to the study by Durbin and Teske[60]. A total of 249 sequences were aligned with MAFFT L-INS-i (v7.012b)[61], trimmed with TrimAl (v1.4)[62], and subjected to a maximum-likelihood phylogeny analysis using RAxML (v8.0.22)[63] (GTRGAMMA model of nucleotide substitution and 100 bootstraps). The resulting tree was imported into iTOL online[64] to collapse major clades.

Phylogenetic analysis of DSAG-related OTU's. All 16S rRNA gene sequences classified as DSAG by Jørgensen *et al.*[52] were used as queries in a BLAST search ($E < 10^{-5}$, identity > 83%) against all archaeal entries in the SILVA database (release 119) that met the following criteria: sequence length > 900 bp, alignment identity > 70, alignment quality > 75 and pintail quality > 75 and the quality of recovered sequences was checked (for example, using 'cut-head' and 'cut-tail' information). The number of sequences in the data set was reduced while keeping maximum diversity as follows. First, the retained 16S rRNA sequences were aligned with SINA (v1.2.11)[65], using all archaea in the SILVA database as reference. The alignment was manually curated with Seaview (v4)[66]. Upon removal of gaps, sequences were used to create OTUs with UCLUST (v1.2.22)[67] (94% identity cut-off and the '–optimal' option). All sequences that corresponded to OTU seeds were selected to represent full DSAG genetic diversity and, upon adding archaeal outgroup sequences and the single amplicon OTU, classified as DSAG, the final data set was aligned with SINA (v1.2.11) as described above, trimmed with TrimAl (v1.4) (gap threshold of 50%) and subjected to RAxML phylogenetic analyses (v7.2.8; GTRGAMMA substitution model, 100 rapid bootstraps). All internal branches with ≤40 bootstrap support were collapsed with Newick-Utilities (v1.6)[68]. The resulting tree was then imported into iTOL online[64] to collapse major clades.

Metagenome sequencing and assembly.

Library preparation and shotgun sequencing. Nextera libraries (Illumina) were prepared according to the manufacturer's instructions, using unamplified LCGC14 (20 ng) and amplified LCGC14AMP (50 ng) as input DNA. Since less starting material was used for the generation of the unamplified library, a total of eight amplification cycles were used in the PCR step during which the Illumina barcodes and adapters (NextEra Index kit) were fused, rather than the default five cycles. The LCGC14 and LCGC14AMP NextEra libraries were sequenced with three and two lanes, respectively, of HiSeq2500 (Illumina), using rapid mode setting, generating two 150-bp paired reads. These runs yielded 8.6 Gbp and 56.6 Gbp of data with an average insert size of 620 and 350 bp for the LCGC14 and LCGC14AMP NextEra libraries, respectively.

Read preprocessing. SeqPrep (v.b5efabc5f7, https://github.com/jstjohn/SeqPrep) was used to merge overlapping paired-end reads and to trim adapters, with default settings. Merged reads and non-merged pairs were trimmed with Sickle (v.1.210, https://github.com/najoshi/sickle), using "se" and "pe" options, respectively, and default settings.

Metagenomic assembly. Pre-processed paired-end reads and single reads were assembled with SPAdes v. 3.0.0[69] in single-cell mode, to take into account the widely varying coverage of metagenomics contigs as well as to try to assemble contigs with low coverage. The read correction tool was turned on and kmers 21, 33, 55 and 77 were used. Mismatch correction was not performed on the LCGC14AMP data set. Contigs shorter than 1 kbp were discarded.

Gene predictions. Protein coding genes (CDS) were identified with prodigal v. 2.60[70], using the 'meta' option for metagenomes. Ribosomal RNA (rRNA) genes were called with rnammer v.1.2[71], using the archaeal model and searching for all three rRNA subunits. Transfer RNA genes (tRNA) were identified with tRNAscan-SE v.1.23[72], using the '–G' option for metagenomes and '–A' option for the Lokiarchaeum composite genome (see subsequent paragraphs). For the latter, the analysis was also run with SPLITSX (no version number available; source code downloaded on 14 August 2014)[73] to detect tRNA genes that are split or that have multiple introns.

Protein clustering. Archaea-specific clusters of orthologous genes (arCOGs)[74], based on 120 archaeal proteomes (hereafter called arCOGs2012), were extended with proteomes from 45 recently sequenced organisms, including 31 single-cell amplified genomes (SAGs) (Supplementary Table 1). First, existing arCOGs were attributed to the new proteomes: protein sequences in each of the 10,323 arCOGs2012 were aligned with MAFFT L-INS-i v.7.130b[61]. Each alignment was used as a query (-in_msa) to search the new proteomes using PSI-BLAST[55], ignoring the master sequence, using 10^{-4} as an *E*-value cut-off, fixing the database size to 10^8, gathering at most 1,000 sequences, and not using composition-based statistics. Hits were then sorted per subject protein and, for each subject, the highest-scoring query alignment was deemed the main arCOG. Whenever applicable, the next-highest, non-overlapping query alignment was deemed the secondary arCOG. Second, proteins without arCOG attribution (singletons) in both the original and extended set of proteomes were gathered, and new arCOGs (arCOGs2014) were created from symmetrical best hits, using the tools available in COG software suite, release 201204 (ref. 75). PSI-BLAST searches were performed according to the COG software instructions. Lineage-specific expansions were identified with COGlse, using a job-description file containing all possible pairs of organisms that do not belong to the same phylum.

COGtriangles was run with default settings, and yielded 3,570 new arCOGs. Of the 325,405 proteins in the combined data sets (165 proteomes), 29,249 (9%) had no arCOG attribution.

Attribution of arCOGs to metagenomes or composite genomes in this study was performed with PSI-BLAST as described above, using the arCOGs2014 as queries.

Phylogenetic analyses of 'taxonomic marker' proteins for binning and concatenated protein trees.

Phylogenetic inference. Maximum-likelihood phylogenies were inferred with RAxML 8.0.9[63], calculating 100 non-parametric bootstraps. PROTGAMMALG and GTRGAMMA were used for amino acid and nucleotide alignments, respectively, unless otherwise stated. Bayesian inference phylogenies were calculated with PhyloBayes MPI 1.5a[76], using the CAT model and a GTR substitution matrix. Four chains were run, and runs were checked for convergence. Whenever convergence was not reached, the topology of individual chains was compared. Consensus trees were obtained with bpcomp, using all four chains and a burn-in of at least half the generations. To add bootstrap support values to the Bayesian phylogenies, sumtrees.py (DendroPy package[77]) was used, with default settings, taking the Bayesian inference tree as a guide tree and the 100 bootstraps as input. For concatenated phylogenies, amino-acid sequences were aligned again with MAFFT L-INS-i individually for each cluster. Positions with >50% gaps were trimmed and alignments were concatenated.

Amino acid bias filtering. To assess the effect of amino acid bias on the phylogenies, a χ^2 filtering analysis was performed on the concatenated alignment. For a complete description, see refs 78 and 79. In brief, a global χ^2 score is calculated for the concatenated alignment, by summing, for each amino acid and each sequence, the normalized squared difference between the expected and observed frequency of the amino acid in this particular sequence and its frequency expected from the whole alignment. Each position in the alignment is individually trimmed and the difference ($\Delta\chi^2$) between the global χ^2 score and the χ^2 score calculated on the trimmed alignment provides an estimation of the relative contribution of each position to the global amino acid composition heterogeneity. Positions are then ranked by their $\Delta\chi^2$ values, and the most or least biased sites up to a threshold are removed.

Tree topology tests. To compare how well different trees explained the aligned sequence data, approximately unbiased tests[80] were performed on concatenated as well as single-gene alignments. Two maximum-likelihood hypothesis trees were tested against the alignments. The first one, showing Lokiarchaeota grouping with eukaryotes, was obtained from the concatenation of 36 markers, shown in Fig. 2b. The second was obtained from the concatenation of the 21 ribosomal proteins present in the previous set, and shows Korarchaeota grouping with eukaryotes. For individual gene trees, the taxa missing in the alignment were also pruned from the hypothesis trees using the utility nw_prune from the Newick Utilities package[68]. For each alignment tested, per-site maximum likelihood was calculated for both hypothesis trees with RaxML 8.0.9, using the option '–f G', and the PROTGAMMALG model. CONSEL 0.20[81] was then used to perform approximately unbiased tests, using default settings.

Identification of taxonomic markers. A reference set of 59 highly conserved, low- or single-copy genes were used both as taxonomic markers in the binning process and for concatenated phylogenies (Supplementary Table 2). Fifty-seven of these, which were shown to be prone to very few or no horizontal gene transfers were taken from ref. 79. Two further arCOGs (arCOG04256 and arCOG04267, subunits A″ and A′ of the DNA-directed RNA polymerase, respectively) were added to the set (see Supplementary Information and Supplementary Table 2 for a list over which arCOG is included in each phylogeny).

Unless otherwise stated, all trees included the same set of 101 reference genomes: 58 archaeal genomes selected[79] from the 120 analysed by Wolf *et al.*[74]; 21 selected from the 45 newly sequenced organisms that were also used for clustering, some of them already analysed in Guy *et al.*[82]; two groups of three closely related SAGs were pooled to provide more complete proteomes; ten bacteria and ten eukaryotes, as in Guy *et al.*[82] (Supplementary Table 1). To remove paralogues and obtain sets with at most one homologue per genome, members of each of the selected arCOGs were aligned with MAFFT L-INS-i and a maximum-likelihood phylogeny was inferred with RAxML, under a PROTCATLG model with 100 slow bootstraps. Previously removed paralogues[79] were not included. Trees were then visually inspected and paralogues removed using the same guidelines as in ref. 79. This set, including at most one copy of each of the 59 reference arCOGs in 101 genomes, is hereafter referred to as '59ref'.

Binning.

Training set. After arCOG attribution (see above), genes from LCGC14AMP belonging to the respective arCOGs were added to the 59ref set. Sequences were aligned and individual trees were built for each arCOG, as described above. Trees were then visually inspected and sequences from LCGC14AMP were classified in the following categories: Lokiarchaeum, Loki2/3 (distant Lokiarchaeaum-related clades), Thaumarchaeota, DPANN, Diapherotrites, Mimivirus, Bacteria or unknown. Classification was based on phylogenetic placement. In some cases where the phylogenetic placement was inconclusive, presence on the same contig of another gene already classified was used to aid classification. The fact that Lokiarchaeum is the only clade for which four to six distinct but closely related strains are present in LCGC14AMP greatly aided classification. In a minority of cases, some genes were classified in a category but marked as 'putative', as their attribution was slightly ambiguous.

Quality control of the training set. Contigs containing markers classified in the first six categories mentioned above were extracted from the assembly, and their tetranucleotide frequencies (TNF) were calculated. To then assess the reliability of the classification, linear discriminant analysis (LDA) was performed in R[83] with package MASS[84], using GC content and TNF as input data: half of the contigs belonging to each of the six selected categories were randomly selected (excluding the contigs marked as 'putative'), and used to calculate LDA (function 'lda' in MASS) (Supplementary Fig. 4). Based on this, classification was predicted using the MASS function predict.lda. Incorrect predictions (that is, when the prediction based on LDA was not congruent with the classification based on the phylogenetic trees) were recorded. The procedure was repeated 100 times, and contigs that were attributed to the wrong category 30 times or more were manually reviewed and eventually discarded from the training set (Supplementary Fig. 4a). Contigs marked as putative were attributed to the category if the prediction was congruent with the putative classification 90 times or more, or discarded otherwise. A further cycle of LDA calculation and prediction was performed, with no contigs classified as 'putative' this time (Supplementary Fig. 4b). To further investigate the robustness of the method, we randomized the categories of the input and performed the same LDA calculation and prediction as above, and assessed the number of incorrect predictions in each case (Supplementary Fig. 4c). This test confirmed that classifications based on trees were generally congruent with predictions based on LDA, significantly more often than just by chance (Supplementary Fig. 4d–f). The final set of contigs was used as a training set for phymmBL[85](see below), and comprised 839 kbp for Lokiarchaeum, 544 kbp for Loki2/3, 521 kbp for Thaumarchaeota, 646 kbp for DPANN, 43 kbp for Diapherotrites and 21 kbp for Mimivirus.

Binning using PhymmBL. PhymmBL version 4.0[85] was run separately for binning the contigs larger than 1 kbp from both LCGC14AMP and LCGC14. As training sets, all prokaryotic genomes published in GenBank (retrieved on 2014-03-04, 2716 genomes) were complemented with the 60 newly sequenced genomes used to constitute the arCOG set that was absent from GenBank (Supplementary Table 1), and the six training sets (Lokiarchaeum, Loki2/3, Thaumarchaeota, DPANN, Diapherotrites and Mimivirus) obtained from LCGC14AMP as described above.

Reassembly of Lokiarchaeum bin. In the LCGC14 assembly, 3,165 contigs (18.6 Mbp in total) were predicted to belong to the Lokiarchaeum genus, indicating a large degree of microdiversity. In order to reduce redundancy, contig sets were constituted, with increasing low-coverage cut-offs (from 1 to 100×, with a 1× increment). The completeness and redundancy of each set was then estimated using the micomplete script (manuscript in preparation). In brief, micomplete bases its predictions on the presence or absence of a set of single-copy pan-orthologs, in this case 162 markers defined in ref. 86. To avoid overemphasizing the presence of markers that are often very close to each other (for example, ribosomal proteins), each marker receives a weight coefficient based on the distance between this marker and its closest neighbours both upstream and downstream, averaged over a representative set of 70 Archaea (set described in ref. 79). Completeness is the fraction of weighted markers present, and is thus constrained between 0 (no marker present) and 1 (all markers present). Redundancy is calculated as the total number of copies of weighted markers present divided by the number of weighted markers present, and is thus always greater than one, where one would mean that all markers present are single copy. These two numbers were calculated for each contig set, and a cut-off of 24× represented the best compromise between completeness (0.89) and redundancy (1.67) (Loki24× set, Supplementary Fig. 5a).

To obtain a better assembly with longer contigs with only reads from Lokiarchaeum, reads belonging to Lokiarchaeum contigs were reassembled as follows. Reads from the LCGC14 data set, corrected by SPAdes, were mapped against the whole LCGC14 assembly with bwa-mem[87], and reads that matched contigs in the Loki24× set were extracted. For paired-end reads, both reads were retained if at least one read matched the Loki24× set. These extracted reads were assembled with SPAdes as above, but without the single-cell mode and without read correction. Again, completeness and coverage were assessed for sets of contigs with increasing low-coverage cut-offs, and a threshold of 20× coverage was found to give the best compromise between completeness (0.92) and redundancy (1.44) (Supplementary Fig. 5b). The selected 504 contigs, hereafter referred to as

'Lokiarchaeum', represented 5.14 Mbp of sequence. The N50 and N90 of this assembly were 15.4 and 5 kbp, respectively.

Annotation and contamination assessment of Lokiarchaeum genome bin. Annotation of all predicted open reading frames of the Lokiarchaeum genome bin was done using prokka[88], using a concatenation of the three kingdom-specific protein databases shipped with prokka as the main database, predicting tRNA and rRNA as above. Furthermore, proteins were compared to sequences in NCBI's non-redundant database and RefSeq using BLAST[55] and results were inspected using MEGAN[89]. Additionally, an InterProScan 5[90] (which integrates a collection of protein signature databases such as BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther, Gene3D, Phobius and Coils) was performed and the genome was viewed and analysed in MAGE[91]. Selected genes of interest for the evolution of the eukaryotic cell and/or subjected to phylogenetic analyses were checked manually and annotated according to their protein domains/signatures based on PSI-BLAST[55] results, arCOG attributions (Supplementary Tables 6–10) as well as protein structure predictions using Phyre2[92]. To check for the presence of particular genes of interest, such as specific eukaryotic ribosomal proteins, or eukaryotic ribosomal protein L41e which has been detected in several Euryarchaeota[93], existing alignments from arCOGs and/or KOGs were downloaded from eggNOG[94] and used in PSI-BLAST searches as query against the Lokiarchaeal composite genome.

Several controls were performed to confirm the absence of obvious contaminants in the final Lokiarchaeum bin. Most importantly, all contigs containing ESPs discussed in the manuscript were manually inspected to verify that these actually belong to Lokiarchaeum, by: (1) inspecting neighbouring genes for the presence of archaeal markers; (2) by querying all proteins present on contigs containing ESPs against the LCGC17 metagenome to check whether highly similar homologues could be found several times in the sample (generally between 3–7 copies) accounting for the different, highly related Lokiarchaeota strains; and (3) by querying the same proteins against environmental metagenomes publicly available at NCBI, controlling that most of them had highly similar homologues in an ocean sediment metagenome[95], but not in any other metagenome. This last check was based on our finding that all ESPs of Lokiarchaeum had highly similar homologues in this marine sediment metagenome (for example, up to 98% for Lokiactins) indicating that closely related genomes of members of Lokiarchaeota are present, which is in accordance with the finding that DSAG represents an abundant group in these sub-seafloor sediments[96].

Finally, proteins comprising informational processing machineries were also investigated using MEGAN[89]. The absence of bacterial informational processing proteins indicated that there is no bacterial contamination in the final bin (see Supplementary Discussion 3).

Identification of taxonomic markers in the bins. For Lokiarchaeum, arCOG attribution was performed as described above, and taxonomic markers were identified by their arCOG attribution. Whenever there were two copies of the same marker, the copy located on the contig with the highest coverage was selected.

For Loki2/3, the category had two copies of 19 out of 36 markers present, with divergent phylogenetic placement. A clear GC content difference could also be observed between the copies, and, with a single exception, the two sets of copies were not overlapping (Supplementary Fig. 6). The exception was discarded and the remaining two-copy markers were divided into two bins, Loki2 (high GC, ranging between 32.2–37.3%) and Loki3 (low GC, ranging between 27.7–30.7%). Single-copy markers with a GC content falling into the range of either of Loki2 or Loki3 were attributed to the corresponding bin, the other copies were discarded. Loki2 (high GC, average 32.8%) consisted of 21 markers and Loki3 (low GC, average 29.9%) of 34 markers.

Taxonomic affiliation of the Lokiarchaeum proteome. To estimate how Lokiarchaeum relates to its closest relatives, its proteome was aligned to NCBI's non-redundant database using blastp, with an *E* threshold of 0.001. To provide a way to compare results, the complete proteomes of '*Candidatus* Korarchaeum cryptofilum' OPF8, '*Candidatus* Caldiarchaeum subterraneum' and the incomplete proteome of SCGC AB-539-E09, sole representative of the Miscellaneous Crenarchaeotal group (MCG) were similarly analysed. The results of the blasts were filtered to remove self-hits and hits to organisms belonging to the same phylum. In the case of the MCG representative, only self-hits were removed. Filtered results were then analysed with MEGAN 5.4.0. Last common ancestor parameters were set as follows: Min Score, 50; Max Expected, 0.01; Top Percent, 5; Min Support, 1; Min Complexity, 0.0. For each result, branches were uncollapsed at the level below super-kingdom. Profiles were compared using Absolute counts, and the results were exported and further analysed in R. Categories to which less than 100 hits were attributed in Lokiarchaeum were grouped under the 'Other Archaea' or 'Other Bacteria' categories. Hits to 'root', viruses, unclassified sequences and hits not assigned were grouped under the 'Other' category. Results are shown in Fig. 2b. Using the same parameters, functional COG categories were assigned to the Lokiarchaeal proteome to get insights into the functional and taxonomic affiliation of the Lokiarchaeal proteome (Supplementary Fig. 15).

Phylogenetic analyses of selected eukaryotic signature proteins (ESPs). *Selection of ESCRT-III homologues.* For the ESCRT-III phylogeny, eukaryotic ESCRT-III homologues as described in Makarova *et al.*[97] (comprising the families Vps60/Vps20/Vps32 and Vps46/2/24), as well as archaeal ESCRT-III homologues belonging to arCOG00452, arCOG00453 and arCOG00454 families present in Crenarchaeota, Thaumarchaeota and Aigarchaeota were extracted from GenBank. The more distantly related SNF7-like arCOG families (arCOG09747, arCOG09749 and arCOG07402)[97] present in a few euryarchaeal species were not included in the alignment. Subsequently, respective arCOGs were retrieved from both the LCGC14AMP metagenome and Lokiarchaeum final bin (see section on arCOG attribution). The ESCRT operon present on a Loki2/3 contig revealed the presence of an additional ESCRT-III homologue (most similar to eukaryotic Vps20/32/60 sequences), which was not attributed to an archaeal COG. This homologue was used as an additional query to retrieve highly similar sequences from the LCGC14AMP metagenome as well as the Lokiarchaeum final bin using blastp. Finally, each of the two different SNF7-family proteins, which are part of the ESCRT operons of Lokiarchaeum and Loki2/3, respectively, were used as queries to search published metagenomes (NCBI) with blastp. Highly similar sequences (coverage $>70\%$; identity $>40\%$) were retrieved and included in the phylogeny as well.

Selection of Vps4 homologues. Archaeal sequences assigned to arCOG01307 (cell division ATPase of the AAA+ class, ESCRT system component) as well as eukaryotic Vps4 homologues, including a few proteins of the cdc48 subfamily, were retrieved from GenBank. The latter protein family served as outgroup, as described in Makarova *et al.*[97] Sequences assigned to arCOG01307 were also extracted from LCGC14AMP metagenome as well as from the Lokiarchaeum bin, and sequences highly similar to the Vps4 of Lokiarchaeum were retrieved from published metagenomes (coverage $>60\%$; identity $>50\%$). The LCGC14AMP metagenome contained a large amount of sequences assigned to arCOG01307, including hits to Vps4 homologues of Thaumarchaeota. However, ATPases that, based on phylogenetic analyses, turned out to be unrelated to Vps4 were removed from the analysis. Based on the initial phylogeny that included all of these sequences, only those LCGC14AMP Vps4 homologues that clustered with the Vps4 homologue of the Lokiarchaeum bin were selected to avoid the inclusion of false positives.

Selection of EAP30-domain (Vps22/36-like) and Vps25 homologues. EAP30 and Vps25 homologues have so far not been detected in Archaea and thus the respective sequences present in Lokiarchaeum (Extended Data Table 1 and Supplementary Table 6) have not been assigned to an arCOG family. Thus, only Lokiarchaeum homologues, as well as selected representative eukaryotic sequences spanning the eukaryotic diversity that were retrieved from the GenBank database were included in these phylogenetic reconstructions. Putative EAP30- and Vps25-like homologues were discovered in the Lokiarchaeum genome since they are part of the ESCRT operon present on contig119. These sequences were used as queries to also retrieve homologues from the LCGC14AMP metagenome (*E* cut-off, 0.1; *q* coverage, 85) as well as from metagenomes deposited at NCBI.

Selection of small GTPase family homologues (IPR006689 and IPR001806). The investigation of the Lokiarchaeum proteome revealed large numbers of proteins homologous to small GTPases of the Ras and Arf families. In order to reliably identify all putative small GTPases in the Lokiarchaeum bin, an InterPro scan[90,98] was performed and all proteins assigned to IPR006689 (Ras type of small GTPases) and IPR001806 (Arf/Sar type of small GTPases) were extracted. Subsequently, archaeal reference sequences belonging to these IPR families were retrieved from GenBank. Eukaryotic and bacterial reference sequences were selected based on a previous study by Dong *et al.*[99] that investigated the phylogenetic relationships of members of the Ras superfamily. Due to the large number of GTPase homologues in the Lokiarchaeum bin, and the difficulty assigning these proteins to a particular taxon, it was decided not to analyse all GTPase homologues present in metagenomes. Upon inspection of the MAFFT L-INS-i alignment, partial sequences and extremely divergent homologues were removed.

Selection of actin homologues. So far, the only actin-related proteins detected in a few members of the archaea belong to arCOG05583 and have been referred to as crenactins[100]. Three proteins encoded by the Lokiarchaeum genome were assigned to this arCOG, and a blastp search against RefSeq revealed that these proteins are more closely related to bona fide actins of Eukaryotes than to archaeal crenactins. In order to identify additional full-length actin homologues, blastp (*E*-value cut-off $<10^{-10}$) searches were performed against the Lokiarchaeum genome as well as the LCGC14AMP metagenome, using this Lokiarchaeum actin

homologue as query. Finally, a total of five and 42 full-length (>180 amino acids) actin-related proteins were retrieved from the Lokiarchaeum bin and from the LCGC14AMP metagenome, respectively. These sequences were merged with the archaeal protein sequences belonging to arCOG05583 as well as with major eukaryotic actin families (actins and ARP1–3 (refs 101, 102)). We also assessed the phylogenetic position of the bacterial actin-related protein (BARP) of the bacterium *Haliangium ochraceum*[103] in light of the new Lokiarchaeal actin homologues, and concluded that the *Haliangium* BARP was most likely acquired via horizontal gene transfer from eukaryotes.

Phylogenetic reconstructions. For all of these ESPs, the selected sequences were aligned using MAFFT L-INS-i[61] and trimmed with TrimAl[62] to retain only those columns present in at least 50% (for ESCRT-III; Vps4; actin homologues), 40% (EAP30-domain and Vps25 homologues) and 80% (small GTPases) of the sequences. Alignments were visually inspected and manually edited whenever necessary and subsequently subjected to maximum-likelihood phylogenetic analyses using RAxML (8.0.22, PROTGAMMALG) with the slow bootstrap option (100 bootstraps).

Contig maps. The contig maps displayed in Fig. 4a were drawn with the software genoPlotR v.0.8.2 (ref. 104).

51. Jorgensen, S. L. *et al.* Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl Acad. Sci. USA* **109,** E2846–E2855 (2012).
52. Jorgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. Quantitative and phylogenetic study of the deep sea archaeal group in sediments of the Arctic Mid-Ocean spreading ridge. *Front. Microbiol.* **4,** 299 (2013).
53. Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180,** 366–376 (1998).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
55. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
56. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10,** 996–998 (2013).
57. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
58. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
59. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).
60. Durbin, A. M. & Teske, A. Archaea in organic-lean and organic-rich marine subsurface sediments: an environmental gradient reflected in distinct phylogenetic lineages. *Front. Microbiol.* **3,** 168 (2012).
61. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
62. Capella-Gutiérrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
63. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
64. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39,** W475–W478 (2011).
65. Pruesse, E., Peplies, J. & Glockner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28,** 1823–1829 (2012).
66. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27,** 221–224 (2010).
67. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
68. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26,** 1669–1670 (2010).
69. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
70. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
71. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
72. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
73. Sugahara, J. *et al.* SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level. *In Silico Biol.* **6,** 411–418 (2006).
74. Wolf, Y. I., Makarova, K. S., Yutin, N. & Koonin, E. V. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct* **7,** 46 (2012).
75. Kristensen, D. M. *et al.* A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26,** 1481–1487 (2010).
76. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62,** 611–615 (2013).
77. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26,** 1569–1571 (2010).
78. Viklund, J., Ettema, T. J. & Andersson, S. G. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29,** 599–615 (2012).
79. Guy, L., Saw, J. H. & Ettema, T. J. The Archaeal Legacy of Eukaryotes: A Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.* **6,** a016022 (2014).
80. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51,** 492–508 (2002).
81. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
82. Guy, L., Spang, A., Saw, J. H. & Ettema, T. J. 'Geoarchaeote NAG1' is a deeply rooting lineage of the archaeal order Thermoproteales rather than a new phylum. *ISME J.* **8,** 1353–1357 (2014).
83. R Core Team. R: A Language and Environment for Statistical Computing (2014).
84. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* 4th edn (Springer, 2002).
85. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* **6,** 673–676 (2009).
86. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
87. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
88. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
89. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).
90. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
91. Vallenet, D. *et al.* MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.* **34,** 53–65 (2006).
92. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4,** 363–371 (2009).
93. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* **7,** e36972 (2012).
94. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42,** D231–D239 (2014).
95. Kawai, M. *et al.* High frequency of phylogenetically diverse reductive dehalogenase-homologous genes in deep subseafloor sedimentary metagenomes. *Front. Microbiol.* **5,** 80 (2014).
96. Morono, Y., Terada, T., Hoshino, T. & Inagaki, F. Hot-alkaline DNA extraction method for deep-subseafloor archaeal communities. *Appl. Environ. Microbiol.* **80,** 1985–1994 (2014).
97. Makarova, K. S., Yutin, N., Bell, S. D. & Koonin, E. V. Evolution of diverse cell division and vesicle formation systems in Archaea. *Nature Rev. Microbiol.* **8,** 731–741 (2010).
98. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–W120 (2005).
99. Dong, J. H., Wen, J. F. & Tian, H. F. Homologs of eukaryotic Ras superfamily proteins in prokaryotes and their novel phylogenetic correlation with their eukaryotic analogs. *Gene* **396,** 116–124 (2007).
100. Ettema, T. J., Lindas, A. C. & Bernander, R. An actin-based cytoskeleton in archaea. *Mol. Microbiol.* **80,** 1052–1061 (2011).
101. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4,** 9 (2009).
102. Goodson, H. V. & Hawse, W. F. Molecular evolution of the actin family. *J. Cell Sci.* **115,** 2619–2622 (2002).
103. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462,** 1056–1060 (2009).
104. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26,** 2334–2335 (2010).

Extended Data Table 1 | Overview of Lokiarchaeal ESPs

Suggested function	Product	Locus tag	IPR-domains	Comment
Putative ESCRT-III proteins	Vps2/24/46-like protein*	**Lokiarch_37480**	IPR005024 Snf7	More distant homologs also present in several other members of the TACK superphylum.
	Vps20/32/60-like protein*	**Lokiarch_16760**	IPR005024 Snf7	
Putative ESCRT-II proteins	EAP30 domain protein (Vps22/36-like)*	**Lokiarch_37450**	IPR007286 EAP30	Previously not found in Archaea.
	Vps25-like protein*	**Lokiarch_37460**	IPR014041 ESCRT-II complex, Vps25 subunit, N-terminal Winged helix; IPR008570 ESCRT-II complex, Vps25 subunit; IPR011991 Winged helix-turn-helix DNA-binding domain	
Putative ESCRT-I protein	Hypothetical protein with Vps28-like domain†	Lokiarch_10170	IPR007143 Vacuolar protein sorting-associated, Vps28	Vps28 is part of ESCRT-I, potential interacting protein Lokiarch_16740 (see Table S6).
Putative ESCRT-associated protein	Vps4 ATPase*	**Lokiarch_37470**	IPR003959 ATPase, AAA-type, core; IPR027417 P-loop containing nucleoside triphosphate hydrolase; IPR003593 AAA+ ATPase domain; IPR007330 MIT-domain	Also present in other members of the Archaea.
Putative vesicular trafficking machinery associated proteins	Hypothetical proteins vacuolar fusion domain MON1‡	Lokiarch_21780 Lokiarch_01670 Lokiarch_15160	IPR004353 Vacuolar fusion protein MON1	Previously not found in other prokaryotic organisms (see Table S6 and Table S10 for more details)
	Hypothetical proteins with longin-like domains	Lokiarch_01890 Lokiarch_13110 Lokiarch_03280 Lokiarch_22790 Lokiarch_04850	IPR011012 Longin-like domain; IPR010908 Longin domain	
	BAR/IMD domain-like superfamily protein‡	**Lokiarch_46220 Lokiarch_08900**	IPR004148 BAR domain; IPR009602 FAM92 protein	Includes various protein families that bind membranes and detect membrane curvature.
Cell division/ cytoskeleton related proteins	Actin and related proteins*	**Lokiarch_44920 Lokiarch_36250 Lokiarch_10650 Lokiarch_09100 Lokiarch_41030**	IPR004000, Actin-related protein; IPR020902 Actin/actin-like conserved site	Some Cren- Kor- and Aigarchaeota encode crenactins [25] (arCOG05583)
	Hypothetical proteins with gelsolin-like domains‡	**12 proteins, Suppl Table S6**	IPR007122 Villin/Gelsolin; IPR029006 ADF-H/Gelsolin-like domain; IPR007123 Gelsolin-like domain	Previously not found in Archaea. Serve as candidates for potential actin-binding proteins.
	Small GTP-binding domain proteins with Ran-/Ras-/Rab-/Rho- and Arf-domain signatures*§	**92 proteins, see Suppl Table S6**	IPR001806 Small GTPase superfamily; IPR003579 Small GTPase superfamily, Rab type; IPR027417 P-loop containing nucleoside triphosphate hydrolase; IPR020849 Small GTPase superfamily, Ras type; IPR002041 Ran GTPase; IPR003578 Small GTPase superfamily, Rho type; IPR005225 Small GTP-binding protein domain; IPR024156 Small GTPase superfamily, ARF type	Extreme proliferation of small GTP-binding proteins in Lokiarchaeum (92 proteins in composite genome, see Fig. 3b and c); in addition Lokiarchaeum encodes 12 Roadblock/LC7 domain proteins, which might serve as GTPase activating enzyme (see Suppl Table S6).
Ubiquitin modifier system related proteins	Ubiquitin-like proteins‡	**Lokiarch_29280 Lokiarch_29310 Lokiarch_37670**	IPR029071 Ubiquitin-related domain; IPR000626 Ubiquitin-like	Ubiquitin modifier system was previously identified in Aigarchaeota[49]; Canonical E3 ubiquitin ligases are not present in *Caldiarchaeum subterraneum* and Lokiarchaeum. However, both archaeal genomes contain RING-domain proteins‡ that could serve as candidates for E3 ligases, e.g. Lokiarch_34010 (see Suppl. Table S6).
	Putative E1-like ubiquitin activating protein	Lokiarch_15900 Lokiarch_29320	IPR023280 Ubiquitin-like 1 activating enzyme, catalytic cysteine domain; IPR019572 Ubiquitin-activating enzyme (see SOM for more details)	
	Putative E2-like ubiquitin conjugating protein	**Lokiarch_10330 Lokiarch_41760 Lokiarch_29330**	IPR016135 Ubiquitin-conjugating enzyme/RWD-like; IPR000608 Ubiquitin-conjugating enzyme, E2	
	Hypothetical proteins with JAB1/MPN/MOV34 metalloenzyme domain	**Lokiarch_29340 Lokiarch_43590 Lokiarch_26830 Lokiarch_08140**	IPR000555 JAB1/MPN/MOV34 metalloenzyme domain	
Eukaryotic ribosomal protein	Putative homolog of eukaryotic ribosomal protein L22e†	**Lokiarch_30160**	-	Previously not found in Archaea. Best blast hit: gb|EPR78232.1| 60S ribosomal protein L22 [*Spraguea lophii* 42_110] - Expect = 0.21
Oligosaccharyl transferase complex proteins	Ribophorin 1 superfamily protein	**Lokiarch_43710**	IPR007676 Ribophorin I	Previously not found in Archaea
	Putative oligosaccharyl transferase complex, subunit OST3/OST6	**Lokiarch_24040 Lokiarch_25040**	IPR021149 Oligosaccharyl transferase complex, subunit OST3/OST6	Previously not found in Archaea.
	Putative oligosaccharyl transferase STT3 subunit	**Lokiarch_28460**	IPR003674 Oligosaccharyl transferase, STT3 subunit	Homologs also present in some other Archaea.

Locus tags that are highlighted in bold indicate a significant top blast hit of the respective protein of Lokiarchaeum to a eukaryotic sequence (see Supplementary Table 6 for further details).
* Phylogenetic analyses have been performed.
† Alignments shown in Supplementary figures.
‡ Protein domain assignments for these proteins listed in Supplementary Table 10.
§ While most small GTPases encoded by Lokiarchaeum have highest similarity to eukaryotic homologues, approximately 10% are most similar to Archaea and/or Bacteria (see Supplementary Table 6 for more details).

ARTICLE

doi:10.1038/nature14416

Neurons for hunger and thirst transmit a negative-valence teaching signal

J. Nicholas Betley[1]*, Shengjin Xu[1]*, Zhen Fang Huang Cao[1]*, Rong Gong[1], Christopher J. Magnus[1], Yang Yu[1] & Scott M. Sternson[1]

Homeostasis is a biological principle for regulation of essential physiological parameters within a set range. Behavioural responses due to deviation from homeostasis are critical for survival, but motivational processes engaged by physiological need states are incompletely understood. We examined motivational characteristics of two separate neuron populations that regulate energy and fluid homeostasis by using cell-type-specific activity manipulations in mice. We found that starvation-sensitive AGRP neurons exhibit properties consistent with a negative-valence teaching signal. Mice avoided activation of AGRP neurons, indicating that AGRP neuron activity has negative valence. AGRP neuron inhibition conditioned preference for flavours and places. Correspondingly, deep-brain calcium imaging revealed that AGRP neuron activity rapidly reduced in response to food-related cues. Complementary experiments activating thirst-promoting neurons also conditioned avoidance. Therefore, these need-sensing neurons condition preference for environmental cues associated with nutrient or water ingestion, which is learned through reduction of negative-valence signals during restoration of homeostasis.

AGRP neurons are a hypothalamic population that is activated or inhibited by hormonal signals of energy deficit or surfeit, respectively[1]. AGRP neuron ablation or inhibition suppresses feeding[2,3], and activation elicits food consumption and instrumental food-seeking within minutes[2,4,5], indicating that these neurons are an entry point to motivational processes resulting from a homeostatic deficit[6]. Because food preferences and food-seeking behaviours are learned, in part, as a consequence of nutrient intake[7], we investigated the capability of AGRP neurons to directly influence learning in mice.

Multiple learning processes contribute to feeding behaviour[8–10]. Behavioural responses to Pavlovian conditioning are typified by approach or avoidance to cues that have been associated with a reinforcer[8,10], such as learning preference for a nutritive food over a non-nutritive object. Instrumental conditioning is a process by which an animal learns to perform an action that elicits a valued outcome, such as lever pressing for food. Neurons that increase food-seeking and consumption in homeostatic hunger may influence these learning processes in two distinct ways. Approach to cues and performance of actions associated with food ingestion can be strengthened through the intrinsic positive valence of nutritive food[7], which is potentiated during energy deficit[8,11–13] (Extended Data Fig. 1a). Alternatively, preference or performance of actions can be conditioned by reducing states with negative valence[11,14–16] (Extended Data Fig. 1b). For neurons that elevate food consumption, their valence can be distinguished, in the absence of food, by whether an animal learns to prefer cues that are associated with increased or decreased activity of these neurons, respectively. Influential experiments with brain stimulation in the lateral hypothalamus found neurons that elicit food intake, have positive valence, and facilitate both Pavlovian and instrumental learning[17,18]. Although early behavioural theories assumed that homeostatic deficits controlled need-based behaviours by reducing a negative internal state[14], neurons that increase food intake have not been reported to signal negative valence[8,11].

Nevertheless, human subjects report that energy deficit is unpleasant and eating can alleviate this feeling[19–21], which indicates a possible role for a poorly understood negative-valence state. Neuronal systems that elicit food-seeking and also mediate the negative valence associated with a homeostatic hunger state have not been identified. Here, we used cell-type-specific neuronal activity manipulations and deep-brain *in vivo* imaging to determine that hunger-promoting AGRP neurons can influence learning and behaviour through negative-valence states.

AGRP neurons condition flavour preference

To investigate whether elevated AGRP neuron activity transmits a negative-valence signal, we performed flavour preference conditioning using *ad libitum* fed mice expressing channelrhodopsin-2 (ChR2) in AGRP neurons ($AGRP^{ChR2}$) (Fig. 1a–d). $AGRP^{ChR2}$ mice and a control $AGRP^{eGFP}$ group were habituated to consume two differently flavoured non-nutritive gels and were then conditioned by separately consuming one flavour during photostimulation and the other without AGRP neuron activation (Fig. 1e). After conditioning, the preference for the flavour consumed by $AGRP^{ChR2}$ mice during AGRP neuron photostimulation was reduced (Fig. 1f). To check whether AGRP neuron photostimulation elicits a general aversive state analogous to the nausea-inducing agent LiCl, we performed a conditioned taste aversion test by pairing a novel taste (saccharin solution) with subsequent photoactivation of AGRP neurons; there was no resulting aversion to saccharin (Extended Data Fig. 2a–c). Together, these experiments demonstrate that a flavour cue associated with high levels of AGRP neuron activity became less preferred, indicating that these neurons transmit a negative-valence signal, but AGRP neurons do not appear to elicit strong aversion or disgust, consistent with the observation that AGRP neuron activation leads to copious food consumption[2,4].

AGRP neuron activity is normally elevated during energy deficit[22,23]. If AGRP neurons contribute to feeding behaviours through a negative-valence signal, then inhibition of AGRP neurons during food restriction would be expected to facilitate learning (Extended Data Fig. 1b). For cell-type-specific chemogenetic inhibition, *Agrp-IRES-Cre* mice were virally

[1]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147, USA.
*These authors contributed equally to this work.

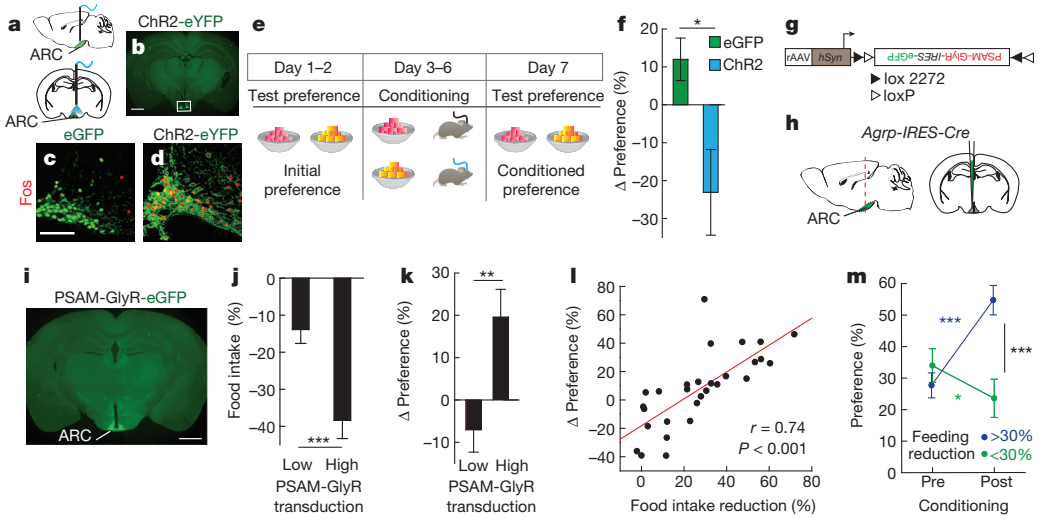


Figure 1 | AGRP neurons condition flavour preference. **a**, Optical fibre position over the arcuate nucleus (ARC). **b**, ChR2–eYFP in AGRP neurons (box). Scale bar, 1 mm. **c**, **d**, Fos immunofluorescence following photostimulation in AGRP^{eGFP} **(c)** or AGRP^{ChR2} **(d)** mice. Scale bar, 100 μm. **e**, Experimental design of conditioned flavour preference assay in *ad libitum* fed AGRP^{ChR2} or AGRP^{eGFP} mice. **f**, Change in preference for flavour paired with light in AGRP^{eGFP} and AGRP^{ChR2} mice (eGFP, $n = 6$; ChR2, $n = 8$). **g**, **h**, Injection of rAAV (**g**) for Cre-dependent expression of PSAM^{L141F}-GlyR-IRES-eGFP in (**h**) AGRP neurons. **i**, Image of virally transduced AGRP neurons showing PSAM^{L141F}-GlyR-IRES-eGFP expression. Scale bar, 1 mm. **j**, Chow food intake reduction for food-restricted mice treated with PSEM^{89S} grouped by transgene transduction efficiency (low, <50%, $n = 13$; high, >50%, $n = 16$). **k**, Change in preference for flavour paired with PSEM^{89S} injection in food-restricted AGRP$^{PSAM\text{-}GlyR}$ mice. **l**, Change in preference correlates with reduction of chow food intake ($n = 29$ mice). **m**, Flavour preference pre- and post-conditioning for mice grouped by post hoc food intake reduction test. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Values are means ± s.e.m. Statistical analysis in Extended Data Table 1.

transduced to express a pharmacologically selective ligand-gated chloride channel, PSAM^{L141F}-GlyR (AGRP$^{PSAM\text{-}GlyR}$ mice)[24] and enhanced green fluorescent protein (eGFP) (Fig. 1g–i and Extended Data Fig. 2d). As previously characterized[24], intraperitoneal (i.p.) administration of the channel's cognate selective synthetic ligand (PSEM^{89S}) inhibited AGRP neuron activity (Extended Data Fig. 2e–g), and reduced food consumption (Fig. 1j and Extended Data Fig. 2h) to an extent that correlated with the transgene transduction efficiency (Extended Data Fig. 2i–n). We measured flavour preference associated with AGRP neuron silencing in food-restricted AGRP$^{PSAM\text{-}GlyR}$ mice (85–90% body weight).

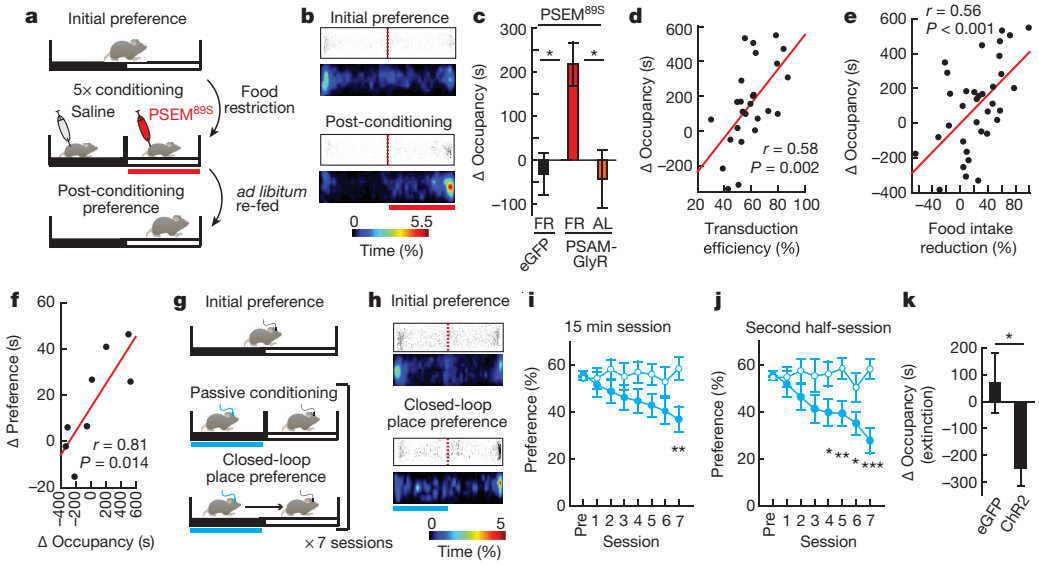


Figure 2 | AGRP neurons condition place preference. **a**, Experimental design of place preference conditioning with chemogenetic silencing. Red bar, chemogenetic silencing side. **b**, For an AGRP$^{PSAM\text{-}GlyR}$ mouse that was food-restricted during conditioning, scatter plot of position and a heat map showing per cent occupancy time. **c**, Change in occupancy time for food-restricted (FR) AGRP^{eGFP} mice ($n = 13$), food-restricted AGRP$^{PSAM\text{-}GlyR}$ mice ($n = 20$) or *ad libitum* (AL) AGRP$^{PSAM\text{-}GlyR}$ mice ($n = 9$) with >50% PSAM^{L141F}-GlyR transduction efficiency after conditioning with PSEM^{89S} injections. **d**, **e**, Change in occupancy time for side paired with PSEM^{89S} is correlated with PSAM^{L141F}-GlyR transduction efficiency ($n = 26$ mice) (**d**) and chow food intake reduction for mice treated with PSEM^{89S} ($n = 35$ mice) (**e**). **f**, Preference shift is correlated for place and flavour conditioning. **g**, Experimental design for place conditioning during optogenetic activation. Blue bar, photostimulated side. **h**, For an AGRP^{ChR2} mouse, scatter plots of position and heat maps showing per cent occupancy time. **i**, **j**, Per cent occupancy time on photostimulated side for AGRP^{eGFP} (open circles, $n = 12$) or AGRP^{ChR2} (filled circles, $n = 12$) mice during 15 min conditioning sessions (two-way repeated measures ANOVA, group: $F_{(1,154)} = 3.0$, $P = 0.097$; session: $F_{(7,154)} = 2.3$, $P = 0.029$; interaction: $F_{(7,154)} = 3.3$, $P = 0.003$) (**i**) and second half of each 15-min session (group: $F_{(1,154)} = 6.4$, $P = 0.019$; session: $F_{(7,154)} = 3.2$, $P = 0.004$; interaction: $F_{(7,154)} = 3.8$, $P < 0.001$) (**j**). **k**, Change in occupancy time on the previously photostimulated side for AGRP^{eGFP} (open circles, $n = 12$) or AGRP^{ChR2} (filled circles, $n = 12$) mice during 1,800 s extinction session. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Values are means ± s.e.m. Statistical analysis in Extended Data Table 1.

Preference increased for the flavour consumed during AGRP neuron inhibition (Fig. 1k), and the change in flavour preference correlated with reduction of chow re-feeding by AGRP neuron silencing (Fig. 1l, m). Therefore inhibiting AGRP neuron activity in energy deficit conditions flavour preference, consistent with suppression of a negative-valence signal.

AGRP neurons condition place preference

We also examined whether AGRP neurons influenced preference for contextual cues independently of ingestive behaviours. First, we performed place preference conditioning by inhibiting AGRP neurons in the absence of food (Fig. 2a). Using a two-sided chamber, food-restricted but not *ad libitum* fed AGRP$^{PSAM\text{-}GlyR}$ mice increased occupancy time in the chamber paired with PSEM^{89S}, whereas control AGRP^{eGFP} food-restricted mice did not shift place preference (Fig. 2a–c). The extent of the shift in occupancy time for food-restricted mice positively correlated with the transduction efficiency of the PSAM^{L141F}-GlyR-IRES-eGFP transgene (Fig. 2d) as well as a post hoc food intake reduction test during AGRP neuron silencing (Fig. 2e). Furthermore, for a subset of mice that were subjected sequentially to both conditioned place preference and flavour preference tests with AGRP neuron silencing, the magnitude of the preference shift was correlated between the two conditioning assays (Fig. 2f). This demonstrates a conditioning process in which reduction of electrical activity in this neuron population during energy deficit reduces food intake and also increases preference for associated contextual and flavour cues.

We next investigated whether the negative-valence properties of AGRP neuron activation influenced conditioned place preference. Although passive conditioning to AGRP neuron stimulation was neither sufficient to reliably change place preference[25] (Δ occupancy time: 28.4 ± 72 s; $P = 0.70$ paired t-test, $n = 10$) nor to oppose cocaine-conditioned place preference (Δ preference, AGRP^{eGFP}/coc: 25.2 ± 4.2%, $n = 6$; AGRP^{ChR2}/coc: 32.7 ± 8.6%, $n = 6$; $P = 0.45$), we sought to determine if the contrast between high and low AGRP neuron activity was learned more effectively. Following passive conditioning, mice were tested each day while freely exploring the conditioning apparatus, and AGRP neuron photostimulation was triggered whenever the mouse entered the side previously exposed to photostimulation (Methods and Fig. 2g). Over the course of multiple sessions, mice showed avoidance of the side paired with AGRP neuron photostimulation (Fig. 2h, i). This effect was more robust for the second half of the closed-loop place preference session (Fig. 2j), which is probably related to the previously reported minutes-long latency of AGRP neuron stimulation to evoke feeding[4]. In a subsequent extinction test in the absence of photostimulation, mice preferred the side that had been associated with cessation of AGRP neuron photostimulation (Fig. 2k and Extended Data Fig. 3a, b). These experiments further indicate that AGRP neurons transmit a negative-valence signal, and discontinuing AGRP neuron photostimulation can condition preference for contextual cues.

The temporal properties and the magnitude of the response to passive place conditioning indicate that AGRP neuron activation is not a strongly aversive stimulus, such as a shock[16], that is capable of eliciting goal-directed instrumental avoidance responses. Indeed, attempts to condition mice to perform an instrumental action (lever-press or nose-poke) to either shut off AGRP neuron photostimulation in well-fed mice or to optogenetically silence AGRP neurons in food-restricted mice were unsuccessful (see Methods and Extended Data Fig. 3c–i). Therefore these experiments indicate that AGRP neuron activity is associated with a negative-valence signal that can mediate Pavlovian learning, but this property does not readily extend to instrumental conditioning.

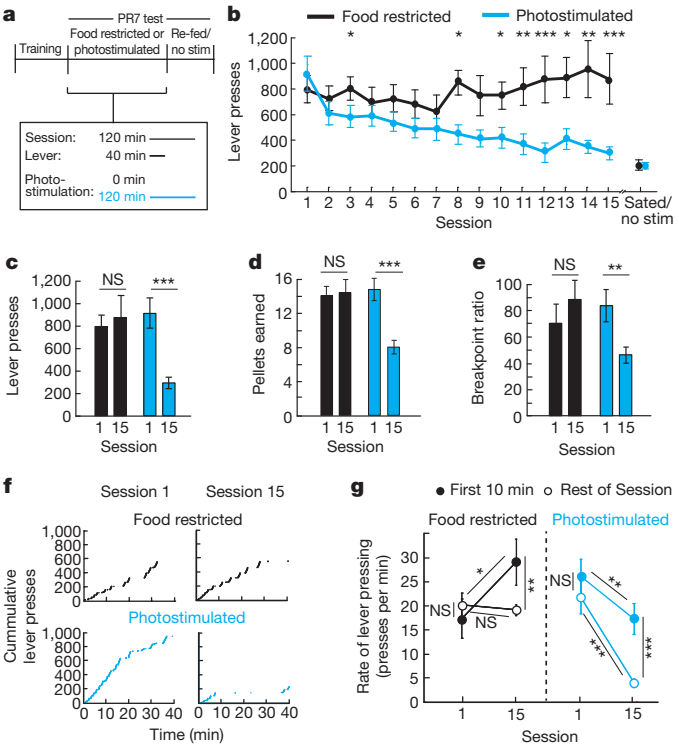


Figure 3 | Modulation of instrumental responding for food. a, Experimental design. AGRP^{ChR2} mice were trained to lever press for food pellets. PR7 reinforcement testing was performed over 15 sessions on two groups: food-restricted (black, $n = 11$) or *ad libitum* fed AGRP neuron photostimulated (cyan, $n = 11$). During test sessions (120 min), levers were available for the first 40 min. The AGRP neuron photostimulated group received intracranial light pulses for the entire 120 min session. **b**, Lever presses in each session during PR7 reinforcement. **c**–**e**, Lever presses (**c**), pellets earned (**d**) and breakpoint ratio (**e**) from first (1) and last (15) PR7 reinforcement test sessions. **f**, Representative traces of cumulative lever pressing during first (1) and last (15) sessions. **g**, Rate of lever pressing during first 10 min of session (low effort reinforcement, filled circles) and rest of session (high-effort reinforcement, open circles) on first (1) and last (15) PR7 session. NS, $P > 0.05$, $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Values are means ± s.e.m. Statistical analysis in Extended Data Table 1.

Modulation of instrumental food-seeking

Although manipulation of AGRP neuron activity does not appear to directly reinforce instrumental responses, AGRP neuron activation leads to vigourous performance of previously learned instrumental food-seeking responses[2,5]. We examined whether AGRP neuron-evoked instrumental responding for food may be sensitive to the negative valence of these neurons. If AGRP neuron activity influences food-seeking behaviours through a negative-valence signal (Extended Data Fig. 1b), then previously reinforced lever-pressing actions would gradually decrease during AGRP neuron photostimulation because nutrient ingestion would not be capable of reducing exogenously elevated AGRP neuron activity. As an alternative hypothesis, if AGRP neuron activity predominantly influenced food-seeking by enhancing the positive valence of food consumption (Extended Data Fig. 1a), then lever-pressing would remain elevated.

We initially compared two groups of mice that were trained to lever-press on a progressive ratio 7 (PR7) food reinforcement schedule under food restriction (Fig. 3a). After learning the contingency between lever-pressing and food delivery, one group that was maintained under food restriction showed steady lever-press responses for 15 sessions. The other group was re-fed *ad libitum* and was tested for instrumental food-seeking during AGRP neuron photostimulation, in which AGRP neuron stimulation was continued after the levers and food access were withdrawn (Fig. 3a). *Ad libitum* fed mice initially

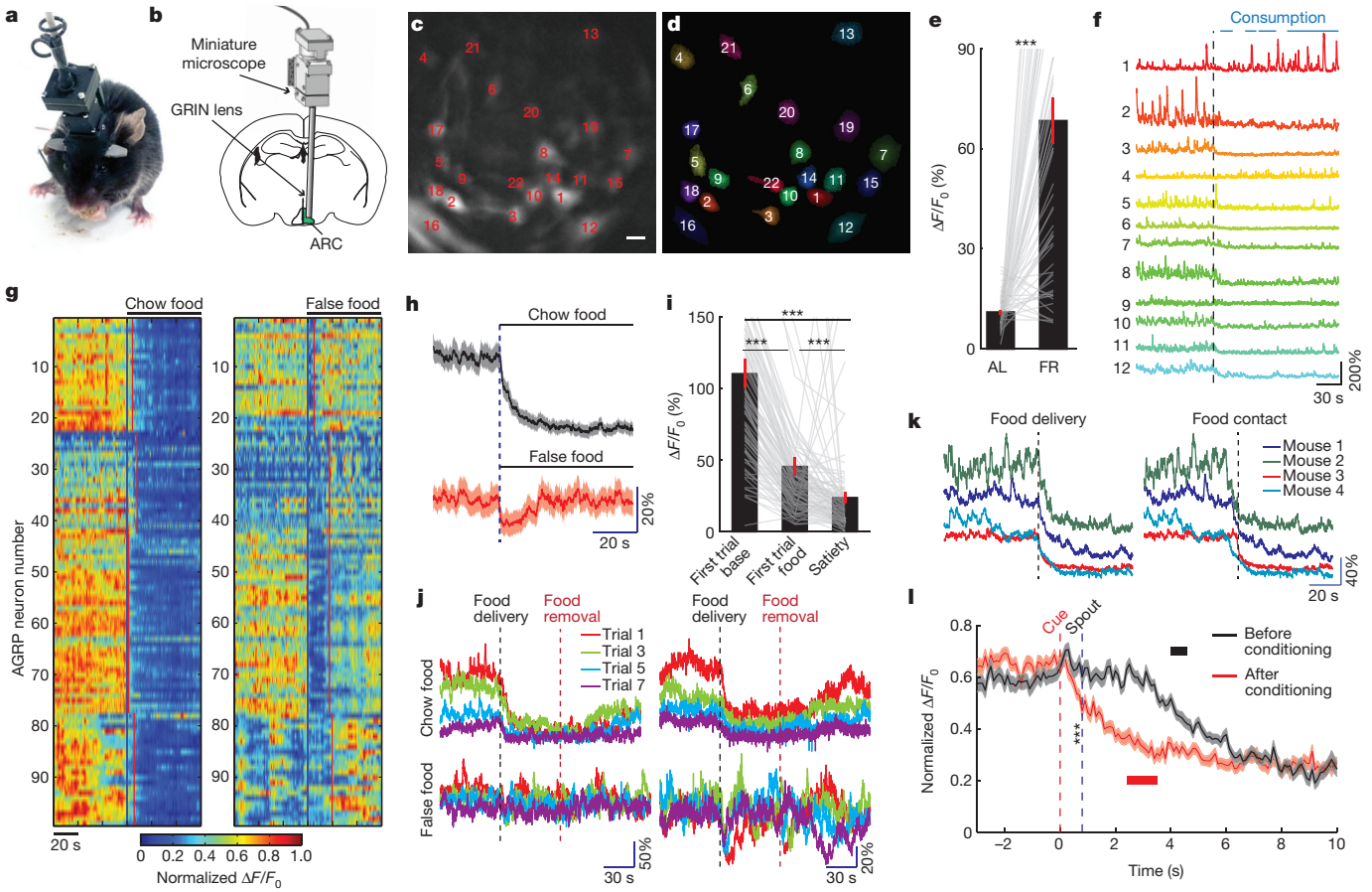


Figure 4 | Food rapidly reduces AGRP neuron activity. **a**, **b**, Configuration for deep-brain calcium imaging from AGRP neurons in freely moving mice. **c**, **d**, Image of AGRP^{GCaMP6f} neurons (**c**) by deep-brain calcium imaging and their region of interest (ROI) spatial filters (**d**) for image analysis. Scale bar, 15 μm. **e**, Change in baseline GCaMP6 fluorescence for neurons in mice under *ad libitum* fed and food-restricted conditions (61 neurons, 4 mice). **f**, From food-restricted mice, GCaMP6f fluorescence traces from subset of individual neurons in **c**, **d**, during chow pellet food consumption. Black line, food delivery. Blue bars, food consumption. **g**, Normalized Ca^{2+} responses of AGRP neurons (99 neurons, 4 food-restricted mice) during exposure to a chow food pellet (left) and a false food pellet (right). Black lines, chow/false food delivery. Red lines, first contact with chow/false food. **h**, Mean calcium responses to chow food and false food aligned to delivery time (99 neurons, 4 food-restricted mice). Shading, s.e.m. **i**, Change in normalized GCaMP6 fluorescence comparing initial baseline activity, first food exposure, and after consuming to satiety (110 neurons, 4 food-restricted mice). **j**, GCaMP6f fluorescence traces from 2 example neurons (2 mice) during short trials of food (top) and false food (bottom) delivery. **k**, Mean GCaMP6 fluorescence responses from individual mice to chow food exposure aligned with food delivery (left) and food contact (right). **l**, Mean GCaMP6 fluorescence responses before (black) and after (red) cued Pavlovian trace conditioning (before, 60 neurons; after, 65 neurons; 3 mice). Black and red bars, range for first lick of liquid food. Shading, s.e.m. ***$P < 0.001$. Values are means ± s.e.m.

responded to AGRP neuron photostimulation with high lever-press rate and food consumption, similar to the food-restricted mice (*t*-test, $P = 0.49$; Fig. 3b). In subsequent sessions, photostimulated *ad libitum* fed AGRP^{ChR2} mice showed a progressive decline in lever presses, pellets consumed, and break point (Fig. 3b–e and Extended Data Fig. 4a–c). Lever-pressing was reduced nearly to the low levels observed without photostimulation (Fig. 3b and Extended Data 4a–c), and pressing at high-effort response ratios was most strongly diminished (Fig. 3f, g). In a separate group, using a shorter photostimulation protocol, lever-pressing was reduced to an intermediate level (Extended Data Fig. 4d–f). We noted an increase in body weight during the multi-session AGRP neuron stimulation protocol, but suppression of lever pressing for food was not due to these long-term metabolic changes (Extended Data Fig. 5). Furthermore, *ad libitum* food intake during AGRP neuron photostimulation was not altered with the extended stimulation protocol, indicating that reduced instrumental food-seeking was not due to food aversion or diminished effectiveness of repeated AGRP neuron photostimulation (Extended Data Fig. 6). Taken together, progressive reduction of AGRP neuron-evoked instrumental food-seeking responses in *ad libitum* mice is consistent with the negative-valence properties of AGRP neurons and indicates reduced value of nutritive food when AGRP neuron activity remains elevated.

Food rapidly inhibits AGRP neurons

AGRP neuron electrical activity manipulations condition learning, but an essential consideration is the correspondence of perturbation studies to the endogenous activity patterns of AGRP neurons during feeding behaviours. To investigate this, we monitored AGRP neuron activity in freely moving mice using deep-brain imaging of genetically encoded calcium indicators through an intracranial gradient index (GRIN) lens with a head-mounted miniature microscope[26] (Fig. 4a, b).

Genetically encoded calcium indicators (GCaMP6f or GCaMP6s[27]) were expressed in AGRP neurons (Fig. 4c, d) and were well tolerated (Extended Data Fig. 7a). Characterization in brain slices revealed sharp increases in calcium activity during burst firing, while changes in tonic firing were detected as a gradual change in the baseline fluorescence (Extended Data Fig. 7b). Characterization *in vivo* by injection of the orexigenic hormone ghrelin substantially increased GCaMP6 brightness and dynamic responses in 81% of AGRP neurons (Extended Data Fig. 7c–e and Supplementary Video 1, 4% of AGRP neurons decreased), which subsequently returned to baseline levels

(population $t_{1/2}$: 19 min, individual neuron $t_{1/2}$ range: 5–46 min, Extended Data Fig. 7f, g). These responses are consistent with previously reported electrical activity changes *ex vivo*[1]. Therefore, imaging AGRP neuron calcium dynamics allows individual neuron activity patterns to be monitored *in vivo*.

We used deep-brain calcium imaging to monitor AGRP neuron activity in food-restricted mice, which was elevated over the *ad libitum* fed condition in 54/61 AGRP neurons (Fig. 4e). Delivery of a mouse chow pellet to food-restricted mice resulted in rapid reduction of GCaMP6 fluorescence during food consumption in 106/110 neurons (96% of neurons, $n = 4$ mice, Fig. 4f–i and Supplementary Video 2); 1/110 neurons increased fluorescence. Removal of the food, after less than 50 mg had been consumed, was followed by a gradual increase in AGRP neuron calcium activity to a level that remained slightly below the initial baseline value (Fig. 4j). In contrast, a false-food pellet (for example, wood block) only transiently reduced AGRP neuron activity, which rapidly recovered after the mouse contacted the object (Fig. 4g, h). Multiple trials with a short exposure to food led to progressive decline in baseline fluorescence that was significantly larger than for the false-food object (Fig. 4j and Extended Data Fig. 7h). These experiments show that in food-restricted mice, baseline AGRP neuron activity was gradually reduced by the consumption of nutritive food, in line with homeostatic regulation, but AGRP neuron activity was also rapidly and strongly suppressed during initiation of food consumption behaviours.

Further analysis of the rapid response to nutritive chow food pellets revealed that calcium activity was reduced before food consumption (Fig. 4k). Moreover, presentation of a visible but inaccessible food pellet, reduced AGRP neuron activity nearly to the same level as during a subsequent food consumption trial (Extended Data Fig. 7i and Supplementary Video 3). These observations demonstrate that AGRP neuron activity is inhibited by food-related cues before nutrients are tasted or consumed. To examine if this rapid AGRP neuron inhibition involves learning, we used Pavlovian trace conditioning to determine the capability of an initially neutral conditioned stimulus to modulate AGRP neuron activity. Initial exposure to a 200 ms auditory and visual compound conditioned stimulus showed a slight increase of mean AGRP neuron calcium activity (Fig. 4l), but GCaMP6 fluorescence was reduced just prior to consumption of a palatable liquid food delivered by subsequent presentation of a lick spout. After repeatedly pairing the conditioned stimulus with food presentation, the conditioned stimulus elicited reduction of AGRP neuron activity (before spout extension, $P < 0.001$, unpaired t-test), and food consumption had little additional effect (Fig. 4l). Therefore, AGRP neurons predictively encode the receipt of nutritive food by rapidly reducing activity, and this process involves learning. Together with neuron silencing and activation experiments (Figs 1–3), these studies demonstrate that endogenous AGRP neuron dynamics during food consumption correspond to the activity manipulations that conditioned preference for flavour and contextual cues. Moreover, the rapid recovery of AGRP neuron activity during false food exposure is expected to reduce preference for non-food objects because AGRP neuron activity signals negative valence.

A virtual thirst state is avoided

Finally, we explored the possibility that a negative-valence signal was used by other homeostatic neurons that mediate a different survival need. To investigate this, we developed an animal model of evoked-thirst by chemogenetic and optogenetic induction of water-seeking and consumption. Prior work has shown the importance of the subfornical organ (SFO) in the brain for mediating water intake[28,29]. Chemogenetic subfornical organ activation selectively elevated consumption of water, but not food, and increased breakpoint for water on a progressive ratio 3 schedule (Extended Data Fig. 8a–i). Elevated drinking was observed by optogenetically activating a subfornical organ neuron subpopulation molecularly defined by expression of nitric oxide synthase 1 (*Nos1*) (Fig. 5a–c). $SFO^{NOS1\text{-}ChR2}$ photostimulation rapidly increased water consumption

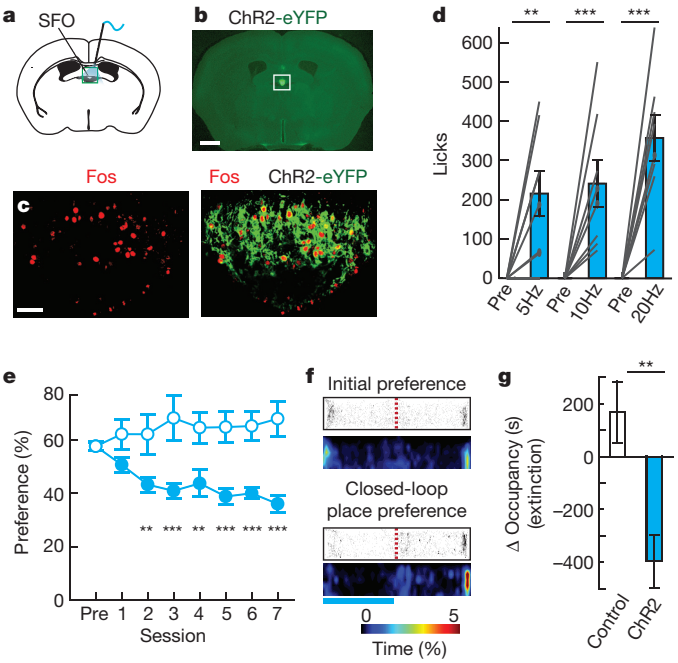


Figure 5 | Virtual dehydration state is avoided. **a**, Optical fibre position over SFO (box). **b**, Expression of ChR2-eYFP in SFO^{NOS1} neurons. Scale bar, 1 mm. **c**, Fos immunofluorescence following photostimulation in $SFO^{NOS1\text{-}ChR2}$ mice. Scale bar, 100 µm. **d**, Water consumption by $SFO^{NOS1\text{-}ChR2}$ mice either before or during photostimulation (1 h) at different frequencies ($n = 8$). **e**, **f**, Closed-loop place preference for $SFO^{NOS1\text{-}ChR2}$ mice (filled circles, $n = 12$) and untransfected controls (open circles, $n = 6$) as in Fig. 3e. Blue bar, photostimulated side. (group: $F_{(1,112)} = 26.2$, $P < 0.001$; session: $F_{(7,112)} = 0.74$, $P = 0.64$; interaction: $F_{(7,112)} = 4.25$, $P < 0.001$). **g**, Change in occupancy time during an extinction session for the photostimulated side for $SFO^{NOS1\text{-}ChR2}$ mice ($n = 12$) and untransfected controls ($n = 6$).*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. Values are means ± s.e.m.

(latency, 20 Hz: 3.8 ± 0.5 min, $n = 8$) at a range of photostimulation frequencies (Fig. 5d) but not food intake (Extended Data Fig. 8j).

To examine the conditioning properties of SFO^{NOS1} neurons, we used the same closed-loop place conditioning protocol as for AGRP neurons (see Fig. 2g). Over the course of 7 sessions, mice showed avoidance of the side paired with SFO^{NOS1} neuron photostimulation (Fig. 5e, f). In a subsequent extinction test in the absence of photostimulation, mice preferred the side associated with cessation of SFO^{NOS1} neuron photostimulation (Fig. 5g). These experiments show that places associated with SFO^{NOS1} neuron activation are avoided, which demonstrates a negative-valence signal that can condition learning from a second homeostatic neuronal cell type with a distinct behavioural function.

Discussion

Physiological need states, in part acting through circulating hormones, lead to elevated electrical activity of specialized need-sensitive neurons, such as AGRP and SFO^{NOS1} neurons. Here we show that these molecularly defined neuron populations signal negative valence. Negative valence physiological states impose a cost on inaction or actions that fail to reduce the need state. Furthermore, through the reduction of negative-valence signals, preference for cues associated with alleviating physiological need states can be learned.

Correspondingly, deep-brain calcium imaging demonstrated that AGRP neuron activity is rapidly inhibited during both food consumption and by cues that predict food. Recent measurement of mean population activity in AGRP neurons also found fast inhibitory dynamics[30], and we show that nearly all AGRP neurons have this property. We also find that rapid reduction of AGRP neuron activity involves the learned association of sensory information with food consumption, highlighting the existence of neural circuit inputs carrying information about conditioned

stimuli. Moreover, sustained reduction of AGRP neuron activity requires nutrient ingestion, consistent with homeostatic regulation, probably involving well-established hormonal control mechanisms.

The valence of increased AGRP neuron activity is opposite to analogous neuronal perturbations in the lateral hypothalamus, which also lead to avid food consumption but exhibit rewarding properties[17,18,31,32]. This may reflect mechanistic differences between homeostatic and hedonic motivation for food, which, for the latter, is primarily distinguished by appetite for highly palatable food even in the absence of a need state[33]. Our experiments, taken together with other studies, indicate that homeostatic need states regulate behaviour through a combination of negative- and positive-valence signals contributing to Pavlovian and instrumental conditioning. Modulation of both processes can be coordinated by hormones such as ghrelin, leptin, and angiotensin[34–36], as well as synaptic inputs[37,38]. Under homeostatic deficit, negative and positive reinforcement processes are expected to operate in a concerted push-pull manner, respectively, to achieve outcomes that have the highest value in that physiological state.

The negative valence of elevated AGRP and SFO^{NOS1} neuron activity is also consistent with human self-reports of negative feelings associated with hunger and thirst arising from homeostatic deficits[20,39]. The behavioural characteristics of AGRP neuron activity in mice parallel some negative emotional aspects of weight-loss in humans, which contribute to low long-term behavioural compliance on weight-loss diets[19,21]. The failure to maintain weight-loss reverses the multifaceted clinical benefits of reduced body weight, such as lessening of diabetes and hypertension symptoms. Our experiments show that AGRP neuron circuits, which are conserved in humans, provide an entry point to investigate the relationship between metabolism and negative emotional states.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 July 2014; accepted 19 March 2015.
Published online 27 April 2015.

1. van den Top, M., Lee, K., Whyment, A. D., Blanks, A. M. & Spanswick, D. Orexigen-sensitive NPY/AgRP pacemaker neurons in the hypothalamic arcuate nucleus. *Nature Neurosci.* **7,** 493–494 (2004).
2. Krashes, M. J. *et al.* Rapid, reversible activation of AgRP neurons drives feeding behavior in mice. *J. Clin. Invest.* **121,** 1424–1428 (2011).
3. Luquet, S., Perez, F. A., Hnasko, T. S. & Palmiter, R. D. NPY/AgRP neurons are essential for feeding in adult mice but can be ablated in neonates. *Science* **310,** 683–685 (2005).
4. Aponte, Y., Atasoy, D. & Sternson, S. M. AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. *Nature Neurosci.* **14,** 351–355 (2011).
5. Atasoy, D., Betley, J. N., Su, H. H. & Sternson, S. M. Deconstruction of a neural circuit for hunger. *Nature* **488,** 172–177 (2012).
6. Sternson, S. M. Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron* **77,** 810–824 (2013).
7. Yiin, Y. M., Ackroff, K. & Sclafani, A. Flavor preferences conditioned by intragastric nutrient infusions in food restricted and free-feeding rats. *Physiol. Behav.* **84,** 217–231 (2005).
8. Dickinson, A. & Balleine, B. in *Stevens' Handbook of Experimental Psychology* Vol. 3 (ed. R. Gallistel) Ch. 12 497–533 (John Wiley & Sons, 2002).
9. Rescorla, R. A. & Solomon, R. L. Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. *Psychol. Rev.* **74,** 151–182 (1967).
10. Mackintosh, N. J. *Conditioning and Associative Learning* (Oxford Univ. Press, 1983).
11. Berridge, K. C. Motivation concepts in behavioral neuroscience. *Physiol. Behav.* **81,** 179–209 (2004).
12. Cabanac, M. Physiological role of pleasure. *Science* **173,** 1103–1107 (1971).
13. Robinson, M. J. & Berridge, K. C. Instant transformation of learned repulsion into motivational "wanting". *Curr. Biol.* **23,** 282–289 (2013).
14. Hull, C. L. *Principles of Behavior* (D. Appleton-Century Co., 1943).
15. Navratilova, E. *et al.* Pain relief produces negative reinforcement through activation of mesolimbic reward-valuation circuitry. *Proc. Natl Acad. Sci. USA* **109,** 20709–20713 (2012).
16. Sidman, M. Avoidance conditioning with brief shock and no exteroceptive warning signal. *Science* **118,** 157–158 (1953).
17. Margules, D. L. & Olds, J. Identical "feeding" and "rewarding" systems in the lateral hypothalamus of rats. *Science* **135,** 374–375 (1962).
18. Jennings, J. H., Rizzi, G., Stamatakis, A. M., Ung, R. L. & Stuber, G. D. The inhibitory circuit architecture of the lateral hypothalamus orchestrates feeding. *Science* **341,** 1517–1521 (2013).
19. Stunkard, A. J. & Rush, J. Dieting and depression reexamined. A critical review of reports of untoward responses during weight reduction for obesity. *Ann. Intern. Med.* **81,** 526–533 (1974).
20. Keys, A., Brozek, J., Henshel, A., Mickelsen, O. & Taylor, H. L. *The Biology of Human Starvation* Vol. II (Univ. of Minnesota Press, 1950).
21. Wadden, T. A., Stunkard, A. J. & Smoller, J. W. Dieting and depression: a methodological study. *J. Consult. Clin. Psychol.* **54,** 869–871 (1986).
22. Takahashi, K. A. & Cone, R. D. Fasting induces a large, leptin-dependent increase in the intrinsic action potential frequency of orexigenic arcuate nucleus neuropeptide Y/Agouti-related protein neurons. *Endocrinology* **146,** 1043–1047 (2005).
23. Betley, J. N., Cao, Z. F., Ritola, K. D. & Sternson, S. M. Parallel, redundant circuit organization for homeostatic control of feeding behavior. *Cell* **155,** 1337–1350 (2013).
24. Magnus, C. J. *et al.* Chemical and genetic engineering of selective ion channel–ligand interactions. *Science* **333,** 1292–1296 (2011).
25. Cravens, R. W. & Renner, K. E. Conditioned hunger. *J. Exp. Psychol.* **81,** 312–316 (1969).
26. Ghosh, K. K. *et al.* Miniaturized integration of a fluorescence microscope. *Nature Methods* **8,** 871–878 (2011).
27. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
28. Johnson, R. F., Beltz, T. G., Thunhorst, R. L. & Johnson, A. K. Investigations on the physiological controls of water and saline intake in C57BL/6 mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **285,** R394–R403 (2003).
29. Oka, Y., Ye, M. & Zuker, C. S. Thirst driving and suppressing signals encoded by distinct neural populations in the brain. *Nature* **520,** 394–352 (2015).
30. Chen, Y., Lin, Y. C., Kuo, T. W. & Knight, Z. A. Sensory detection of food rapidly modulates arcuate feeding circuits. *Cell* **160,** 829–841 (2015).
31. Olds, J. in *Brain Stimulation and Motivation: Research and Commentary* (ed. E.S. Valenstein) 81–99 (Scott, Foresman and Company, 1973).
32. Jennings, J. H. *et al.* Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell* **160,** 516–527 (2015).
33. Saper, C. B., Chou, T. C. & Elmquist, J. K. The need to feed: homeostatic and hedonic control of eating. *Neuron* **36,** 199–211 (2002).
34. Domingos, A. I. *et al.* Leptin regulates the reward value of nutrient. *Nature Neurosci.* **14,** 1562–1568 (2011).
35. Egecioglu, E. *et al.* Ghrelin increases intake of rewarding food in rodents. *Addict. Biol.* **15,** 304–311 (2010).
36. Wang, Q. *et al.* Arcuate AgRP neurons mediate orexigenic and glucoregulatory actions of ghrelin. *Mol. Metabol.* **3,** 64–72 (2014).
37. Yang, Y., Atasoy, D., Su, H. H. & Sternson, S. M. Hunger states switch a flip-flop memory circuit via a synaptic AMPK-dependent positive feedback loop. *Cell* **146,** 992–1003 (2011).
38. Krashes, M. J. *et al.* An excitatory paraventricular nucleus to AgRP neuron circuit that drives hunger. *Nature* **507,** 238–242 (2014).
39. Rolls, B. J. *et al.* Thirst following water deprivation in humans. *Am. J. Physiol.* **239,** R476–R482 (1980).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was funded by the Howard Hughes Medical Institute. Z.F.H.C. was funded by the HHMI Janelia Farm Graduate Scholar program. We thank B. Balleine, M. Schnitzer, N. Ji, A. Lee, Z. Guo for suggestions on experimental design; H. Su for molecular biology; J. Rouchard, S. Lindo, K. Morris, M. McManus for mouse breeding and procedures; M. Copeland for histology; J. Osborne and C. Werner for apparatus design; K. Branson for automated mouse tracking software (Ctrax); J. Dudman, S. Eddy, H. Grill, N. Geary, U. Heberlein for comments on the manuscript.

Author Contributions J.N.B. and S.M.S initiated the project. J.N.B., Z.F.H.C., S.X. and S.M.S. prepared the manuscript with comments from all authors. J.N.B., S.X., Z.F.H.C., R.G., C.J.M. and S.M.S. designed the experiments and analysed the data. J.N.B. and Z.F.H.C. performed conditioning experiments, S.X. performed *in vivo* calcium imaging experiments, R.G. and C.J.M. developed the SFO activation model for evoked water drinking, Y.Y. helped with image registration.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M.S. (sternsons@janelia.hhmi.org).

METHODS

All experimental protocols were conducted according to US National Institutes of Health guidelines for animal research and approved by the Institutional Animal Care and Use Committee at Janelia Farm Research Campus.

Mice. Mice were housed on a 06:00 to 18:00 light cycle with water and mouse chow *ad libitum* (PicoLab Rodent Diet 20, 5053 tablet, TestDiet) unless otherwise noted. Adult male mice (>8 weeks old) were used for experiments. Cre recombinase-expressing lines were used: *Agrp-IRES-Cre* (Jackson Labs Stock 012899, *Agrp*$^{tm1(cre)Lowl}$/J), *Nos1-IRES-Cre* (Jackson Labs Stock 017526, B6.129-*Nos1*$^{tm1(cre)Mgmj}$/J). For channelrhodopsin-2 expression in AGRP neurons, *Agrp-IRES-Cre* mice were crossed with *Ai32: ROSA26-loxStoplox-ChR2-eYFP* (Jackson Labs stock 012569, B6;129S-*Gt(ROSA)26Sor*$^{tm32(CAG-COP4*H134R/EYFP)Hze}$/J). For eGFP in AGRP neurons, *Agrp-IRES-Cre* mice were crossed with ROSA-GNZ: *ROSA26-loxStoplox-GFP-NLS-LacZ* (Jackson Labs stock 008516). For Arch in AGRP neurons, *Agrp-IRES-Cre* mice were crossed with *Ai35d*: *ROSA-CAG-loxStoplox-Arch-GFP-WPRE* (Jackson Labs stock 012735).C57BL/6J mice were from Jackson Labs (stock 000664).

Recombinant adeno-associated viral (rAAV) vectors. The following Cre-dependent viral vectors[40,41] were used in this study: rAAV2/10-*CAG*-FLEX-*rev*-ChR2tdtomato (3e13 Genomic Copies (GC) per ml, University of Pennsylvania vector core), rAAV2/1 and rAAV2/9-*hSyn*-FLEX-*rev*-PSAM^{L141F}-GlyR-IRES-eGFP (1.4e13 and 1.5e13 GC per ml, respectively, Janelia, http://www.addgene.org/Scott_Sternson/), rAAV2/9-*CAG*-FLEX-eGFP (7e12 GC per ml, Penn), rAAV2/1-hSyn-Cre (5.6e12 GC per ml, Janelia), rAAV2/2-*Ef1a*-DIO-hChR2(H134R)-EYFP (6e12 GC per ml, UNC vector core), rAAV2/2-*Ef1a*-DIO-hM3D(Gq)-mCherry (3e12 GC per ml, UNC), AAV2/1-*Syn*-FLEX-GCaMP6f and AAV2/1-*Syn*-FLEX-GCaMP6 s (10^{12} to 10^{13} GC per ml, Janelia). *CAG*, promoter containing a cytomegalovirus enhancer; the promoter, first exon and first intron of the chicken beta actin gene; and the splice acceptor of rabbit beta-globin gene. *Efl1a*, human elongation factor-1 alpha promoter. FLEX, Cre-dependent flip-excision switch. DIO, double-floxed inverted orientation.

Viral injections and optical fibre placement. Viral injections and implantation of ferrule-capped optical fibres (200 μm diameter core, multimode, NA 0.48, ThorLabs) were performed as described previously[23]. Bilateral ARC viral injections in *Agrp-IRES-Cre* mice were made at two depths using the following coordinates: bregma: −1.3 mm; midline: ± 0.3 mm; dorsal surface: −5.95 mm and −5.85 mm (250–500 nl per site). Bilateral SFO viral injections were made at bregma: −0.35 mm; midline: ± 0.6 mm; dorsal surface: −2.45 mm (100–300 nl per site). For deep-brain calcium imaging, unilateral viral injections were made at bregma: −1.46 mm; midline: 0.3 mm; dorsal surface: −6.0 mm and −5.85 mm (100 nl per site).

After 2 to 4 weeks for transgene expression, a ferrule-capped optical fibre was placed for AGRP^{ChR2} mice over the ARC (bregma −1.4 mm; midline: +0.25 mm; dorsal surface 5.6 mm) and for SFO$^{Nos1\text{-}ChR2}$ mice over the SFO (bregma: −0.35 mm; midline: ± 0.6 mm; dorsal surface: −2.3 mm; approach angle: 12°).

Immunohistochemistry. Immunohistochemistry was as described previously[23]. Antibodies: goat anti-AGRP (1:5,000, Neuromics, GT15023), guinea pig anti-RFP (1:25,000, Covance), rabbit anti-Fos (1:5,000, Santa Cruz, SC-52, Lot-C1010), rabbit anti-GFP (1:5,000, Invitrogen, A-11122). Confocal images (Zeiss LSM 510 microscope) were acquired first from Fos immunostained tissue taken from a food deprived mouse, as described previously[23]. These image acquisition settings were maintained for all quantitative Fos analysis (each condition: >50 nuclei from 3 mice, selected blind to the Fos immunofluorescence levels). Transgene transduction efficiency in AGRP neurons was determined as previously described[23] (from >500 AGRP boutons, multiple sections).

Food restriction. Food intake was adjusted to maintain mice at 85–90% of their initial *ad libitum* fed body weight, and food was consumed 18 h before subsequent behavioural assays.

AGRP neuron inhibition *in vivo*. For AGRP neuron silencing, mice expressing PSAM^{L141F}-GlyR in AGRP neurons (AGRP$^{PSAM\text{-}GlyR}$ mice) were injected intraperitoneally (i.p.) with the ligand PSEM^{89S} (30 mg per kg) dissolved in saline.

Suppression of refeeding in food-restricted mice. In food-restricted AGRP$^{PSAM\text{-}GlyR}$ mice two 1 h food intake measurements were performed early in the light period and separated by a day. Saline or PSEM^{89S} (30 mg per kg) were administered 30 min before food was provided, at the time food was provided, and 30 min later (multiple administration of PSEM^{89S} was used due to its rapid clearance[24]).

Suppression of Fos expression by AGRP neuron silencing. In *ad libitum* fed mice, Fos immunofluorescence intensity was measured in AGRP neurons from AGRP$^{PSAM\text{-}GlyR}$ ($n = 3$) and AGRP^{eGFP} ($n = 3$) during the dark period after PSEM^{89S} treatment (30 mg per kg). Fos expression in AGRP neurons is elevated in the dark period in the absence of food, which was removed from the mice at the beginning of the dark period (18:00). PSEM^{89S} (30 mg per kg) was injected 30 min before the onset of the dark period and subsequently every 45 min for 5 h (injection frequency is due to pharmacokinetics of PSEM^{89S})[24]. Mice were deeply anaesthetized and then perfused, and the brain was dissected for immunohistochemical analysis.

Suppression of dark period feeding. Ad libitum fed mice were injected with either saline (test day 1) or PSEM^{89S} (test day 2). At the onset of the dark period, mice were injected again and given free access to chow. PSEM^{89S} (30 mg per kg) or saline was administered every hour until the assay concluded at 22:00; food consumption was recorded.

Photostimulation *in vivo*. Photostimulation was as described previously[23]. Light pulse protocol: 10 ms pulses, 20 Hz (unless otherwise noted) for 1 s, repeated every 4 s.

Conditioned flavour preference. Food-restricted AGRP$^{PSAM\text{-}GlyR}$ mice and AGRP^{eGFP} control mice were acclimatized for four sessions (15 min) to consumption of two non-nutritive gels that were sweetened with sucralose but differed by flavour (orange and strawberry). Prepackaged Hunts sugar-free Juicy Gels were used that contained 0.05 kcal per g (for comparison, 4.1 kcal per g in chow). Consumption during last two sessions was used to determine initial flavour preference (no significant initial group flavour preference, $P = 0.41$, *t*-test). For flavour preference conditioning, mice were given two daily 30 min sessions (repeated over 4 days) with each gel individually, separated by 4 h, with the order of conditioning for the gels inverted each day. The initially preferred flavoured gel was presented paired with saline injection, and the less preferred flavoured gel was paired with injection of PSEM^{89S} (30 mg per kg) (each injection after 5 min of consumption). After conditioning, equal quantities of the two gels were presented (15 min) and the amount of each flavour consumed was recorded. This was repeated the following day with the position of the gel inverted and preferences from the two test sessions were averaged.

To examine conditioned flavour preference learning during AGRP neuron activation, AGRP^{ChR2} mice and AGRP^{eGFP} control mice implanted with a ferrule-capped optical fibre over the ARC were used. Mice were conditioned as above with photostimulation, with the following differences. *Ad libitum* fed mice were used and were acclimatized to consume the two non-nutritive overnight (3 g). For conditioning, in one session, the mouse was presented with 0.3 g of its preferred flavoured gel and after 5 min, intracranial light pulses were applied for an additional 25 min. In the other session, the mouse was presented with 0.3 g of the less preferred flavoured gel, and the mouse was kept in the cage for 30 min without any light applied to the fibre. Mice typically ate the entire 0.3 g during each session. After conditioning, the *ad libitum* fed mice were presented again with equal quantities of the two gels for two 15 min test sessions as described above.

Conditioned taste aversion. AGRP^{ChR2} mice with implanted ferrule-capped optical fibres were used for all groups (LiCl, saline, photostimulation, no photostimulation; all $n = 6$ mice). Mice were placed on water-restriction. After acclimation to sipper tubes, the four groups were allowed to consume a tastant (0.15% saccharin solution, 20 min), and the amount was recorded. For LiCl and saline groups, mice were injected with either LiCl (125 mg per kg) or saline (0.9%) immediately following the exposure to the tastant. After 48 h, consumption of tastant was tested (20 min). On the next day, consumption of water (20 min) was measured. For the photostimulation group, mice received AGRP neuron photostimulation for 120 min immediately following exposure to tastant. The no photostimulation group was tethered with a fibre for the same period of time, but no light pulses were delivered. On the next day, consumption of tastant solution (20 min) was measured, followed by further photostimulation conditioning (total: 4 conditioning and 4 test sessions). The day after the last test session (test 4), consumption of water (20 min) was measured.

Conditioned place preference. A two chambered apparatus was used with visual (black and white sides) and textural cues (black side: plastic grid (3 mm holes) flooring, white side: soft textured side of Kimtech bench-top-protector #7546 Kimberly-Clark). The floor was back-lit (luminance ~100 Lux), and an overhead video camera recorded position (Basler, 3.75 Hz frame rate, gVision software, http://gvision-hhmi.sourceforge.net/). The apparatus was in a sound isolation chamber. *Ad libitum* fed AGRP$^{PSAM\text{-}GlyR}$ mice and AGRP^{eGFP} controls were acclimatized (15 min). The following day, the mouse's position was recorded (1,800 s) and tracked offline using Ctrax[42], and the initial side preference was determined. Mice were then food restricted (FR) to 85–90% of their initial body weight before conditioning sessions. Daily conditioning consisted of two 1,800 s sessions: (1) the initially preferred side of the chamber paired with saline injection; (2) the initially less preferred side paired with PSEM^{89S} (30 mg per kg) injection. After five conditioning days, mice were *ad libitum* re-fed. The following day, mice were given free access to the entire apparatus and their position was tracked. The change in occupancy is the change in time spent on the initially less

preferred side following training, for which positive numbers reflect increased preference for the side on which they were injected with PSEM^{89S}. An additional set of AGRP$^{PSAM\text{-}GlyR}$ mice was tested on the above protocol with the following modification: before the daily conditioning sessions, mice were given free access to mouse chow for 2.5 h, during which time they consumed 2–3 g.

Closed-loop place preference. Conditioning for AGRP^{ChR2} optogenetic neuron photostimulation was performed with a two-step protocol alternating passive conditioning with a closed-loop place preference test. The closed-loop place preference protocol alone was less reliable for conditioning avoidance, likely related to the minutes-long latency[4] of AGRP neuron activation to induce food-seeking and consumption. Therefore, some prior experience with prolonged AGRP neuron activity on one of the sides appears to improve efficacy, although this exposure alone did not significantly alter place preference. Passive conditioning sessions involved separate exposure to each side of the apparatus (1,800 s each) with the mouse tethered to the fibre, where the initially more preferred side was paired with intracranial light pulses to photostimulate AGRP neurons. Passive conditioning was followed by closed-loop place preference testing on the subsequent day, in which mice were allowed free access to both sides of the chamber and photostimulation was applied when the mouse entered the side of the chamber also paired with photostimulation during passive conditioning (photostimulation ceased as soon as the mouse crossed to the other side). After 7 conditioning-days (morning, closed-loop place preference; afternoon, passive conditioning), the mice were given free access to the chamber for 1,800 s without photostimulation (extinction test). AGRP^{eGFP} control mice were treated identically. The same protocol was used for conditioning SFO$^{NOS1\text{-}ChR2}$ mice (Fig. 5).

Cocaine place preference. Conditioning for AGRP^{ChR2} or AGRP^{eGFP} optogenetic neuron photostimulation was performed in combination with a cocaine injection. Three passive conditioning sessions involved separate exposure to each side of the apparatus (1,800 s each) where the initially less preferred side was paired with intracranial light pulses and i.p. cocaine administration (10 mg per kg), and the initially preferred side was paired with saline injection and no photostimulation.

Post hoc assessment of conditioned appetite. After conditioning avoidance of the side associated with AGRP neuron activation, AGRP^{ChR2} mice were subsequently tested for the possibility that AGRP neuron activation might have conditioned elevated appetite in the context associated with prior AGRP neuron activation. *Ad libitum* fed mice in the light period (09:00 to 12:00) were restricted to the side of the chamber previously associated with photostimulation in which they were given free access to food for 1 h. In a subsequent session performed on a different day, the mouse was confined to the other side of the chamber and also given free access to food for 1 h. The side order was counterbalanced, and exposure to food on both sides was in the absence of photostimulation.

Instrumental conditioning. Training and tests were conducted in operant conditioning chambers (Coulbourn Instruments) housed in sound isolation chambers as previously described[23].

Negative reinforcement. Male AGRP^{ChR2} mice implanted with a ferrule-capped optical fibre over the ARC were used if they consumed at least 0.7 g of food (20 mg grain pellets, TestDiet) during photostimulation as a test for proper fibre placement. We adapted a negative reinforcement instrumental conditioning protocol that was previously established for avoidance of optogenetic lateral habenula photostimulation[43]. Conditioning chambers had two nose poke ports, both with backlighting: one active port to stop AGRP neuron photostimulation and one inactive. AGRP^{ChR2} mice were photostimulated and each nose poke resulted in a 20 s pause for light pulses, and a tone and houselight cue were turned on until the laser stimulation returned (session time: 20 min).

A different set of male *ad libitum* fed AGRP^{ChR2} mice were first trained to perform a lever pressing task on a fixed ratio schedule (FR1) in order to obtain food. Training occurred overnight, and a minimum of 50 lever presses was the inclusion criterion. The protocol above was applied where photostimulation pause was contingent on lever pressing.

Instrumental conditioning for AGRP neuron inhibition. AGRP^{Arch} mice (*Agrp-IRES-Cre;Ai35d*) were food restricted to 85–90% of their body weight. Testing was in sound isolation boxes with operant chambers containing one backlit nose poke port. Each nose poke resulted in 60 s of light (561 nm) delivery at a power of 15 mW delivered from fibre tip, in conjunction with onset of tone and houselight cues (session length, 1 h). AGRP neurons were estimated to receive irradiance of $>$15 mW per mm^2 (http://www.openoptogenetics.org).

Progressive ratio 7 food reinforcement schedule. Male AGRP^{ChR2} mice implanted with a ferrule-capped optical fibre over the ARC were used for instrumental conditioning in a lever-press task for food. Mice were used that consumed at least 0.7 g of food during a 1 h photostimulation test. Mice that did not reach this criterion were used for the control (food-restricted) group.

Experiments were conducted in operant conditioning chambers with two retractable levers, one active (delivered food pellets) and one inactive (did not deliver food pellets), at either side of the food hopper, and chambers were housed in sound isolation boxes. Levers were extended at the start of each session. After reaching lever-press criteria for a reward, levers were retracted and a food pellet was delivered. Five seconds after pellet removal from the food hopper, levers were extended again.

Training. Food rewards during training consisted of grain pellets (20 mg) with identical composition to homecage food. Lever-press training was conducted under food restriction, where mice were maintained at ~85% body weight. All mice were trained to lever press for food with a FR1 reinforcement schedule overnight for one night. Mice were trained on daily, 30 min FR1 sessions until reaching learning criteria (earning 18 pellets in a 30 min session for 3 consecutive days). They were then trained with 2 h sessions for two days on a progressive ratio schedule where the required number of presses for each subsequent reward increases by 3 (PR3). Mice were then trained on a PR7 schedule for one day in a 2 h session.

Progressive ratio 7 reinforcement. Following training, mice were tested on PR7 test for 15 consecutive days. PR7 test sessions were 2 h, however, mice were only allowed to lever press for food for the first 40 min of each session. At the start of PR7 testing, food rewards were switched to 20 mg grain pellets with 1% saccharin and grape flavouring (TestDiet) to allow comparison of reinforcing effects of food consumption outside of the testing session (see below). All mice received exposure to and consumed these grape flavoured pellets in their homecages (50 pellets available) the night before the start of tests to limit neophobia.

Food-restricted group: Mice in the food-restricted group were tested while tethered to a dummy fibre to ensure that this tether does not interfere with lever pressing activity or with food consumption.

AGRP neuron stimulation groups: For the AGRP neuron stimulation groups, mice were returned to *ad libitum* food intake for 2 days after training, before initiating testing. Mice were maintained under well-fed conditions. During PR7 test sessions, one group of mice received photostimulation for the whole length of the session (2 h). Within this group, some mice were provided regular chow in their home cages, while others were maintained on the same food used as rewards during the test session for the duration of the experiment. These two subgroups were ultimately combined for statistical analysis because no difference in lever pressing was observed between them. A second group received photostimulation only during the first 40 min of the session, when the mice were allowed to press for food. Mice in both photostimulation groups were also tested for lever pressing in the absence of photostimulation after the 15 day test sessions.

Food restricted with AGRP neuron stimulation group After training, a group of food-restricted mice was tested with AGRP neuron photostimulation during a PR7 food reinforcement schedule with photostimulation for the whole length of the session (2 h). Mice were then returned to *ad libitum* food for 2 days, and retested on a PR7 schedule under well-fed conditions without photostimulation.

No stimulation group After training, mice in the no stimulation group were returned to *ad libitum* food intake for 2 days, before initiating testing. All mice were maintained under well-fed conditions without photostimulation for testing on a PR7 schedule for 16 consecutive days. Mice were tethered to a dummy fibre during testing.

Mice that did not earn at least 5 food rewards on the first day of PR7 test were removed from the experiment (one mouse from the food-restricted group). Breakpoint was defined as the last ratio completed before 5 min passed without earning a reward. For rate of lever pressing analysis, lever presses were divided into two blocks: first 10 min (low-effort work requirement) and rest of session (high-effort work requirement). The first 10 min were chosen as low-effort work requirement conditions as average breakpoint time was greater than 10 min for all groups.

Photostimulation-induced weight gain. We noted an increase in body weight during the multi-session AGRP neuron stimulation protocol (Extended Data Fig. 5a), probably due to a long-acting effect of released AGRP[44] following photostimulation (this is not responsible for acute food consumption under investigation here, which is due to the release of neuropeptide Y and GABA[4,5,45]). To examine whether suppression of lever pressing for food was due to these long-term metabolic changes, we performed control experiments with photostimulation-induced body weight gain separately from interference with negative reinforcement.

AGRP^{ChR2} mice were trained to lever press under food restriction as described above. After training, mice were returned to *ad libitum* food for 2 days before testing began. Food rewards were switched to 20 mg grain pellets with 1% saccharin and grape flavouring (TestDiet) during PR7 testing, mirroring the protocol used in the PR7 reinforcement assay above, and only the differences are described. Mice were then tested under PR7 food reinforcement during photostimulation of AGRP neurons. Next, mice underwent the weight gain induction

period. Weight gain was induced in one group of mice through daily, 2 h photostimulation sessions (22 days) until weight gain matched that of the 2 h photostimulation group in the PR7 test (~28%). A second group did not receive photostimulation, but were tethered to an optic fibre and served as controls for natural weight gain and any potential decline in lever pressing due to the time elapsed between tests. During these sessions, mice did not lever press for food. Mice in both groups were allowed to consume 20 mg grain pellets with 1% saccharin and grape flavouring with the number of rewards each day matched to the average number of rewards on the corresponding session from the 2 h photostimulation group in the PR7 test. After ~28% weight gain in the photostimulated group, well-fed mice were then tested again on a PR7 task during AGRP neuron stimulation.

Repeated daily AGRP neuron-evoked free feeding assay. To assess the consequences of repeated daily AGRP neuron photostimulation sessions on *ad libitum* food intake, mice were tested as in the PR7 experiment, but pellets were freely delivered without levers present.

GRIN lens implantation and baseplate fixation. Mice expressing GCaMP6f or GCaMP6 s in AGRP neurons were anaesthetized using isoflurane, and a rectangle craniotomy (2–3 mm) was made around viral injection coordinates (bregma, −1.46 mm; midline, 0.3 mm). A customized sharp optical fibre (diameter: 0.6 mm) was inserted to the brain to ~250 μm above the ARC. After retraction of the fibre, a gradient index (GRIN) lens (Part ID: GLP-0584; diameter: 0.5 mm, length: 8.2 mm; Inscopix) with a custom GRIN lens-holder was slowly (150 μm per min) implanted. The target depth was determined by observing fluorescent signal through a miniature microscope (nVista HD and HD v2, Inscopix). The GRIN lens was fixed with black dental cement (Lang Dental Manufacturing); then a head bar was fixed with dental cement. A layer of parafilm was covered the top end of the lens. A silicone adhesive (Kwik-Sil; World Precision Instruments) was applied above the parafilm to protect the lens.

Two to four weeks after GRIN lens implantation, awake mice were head-fixed by a head bar holder. A baseplate (Part ID: BPL-1 and Part ID: BPL-2; Inscopix) attached to the miniature microscope was positioned above the GRIN lens. The focal plane was adjusted until neuronal structures and GCaMP6 dynamic responses were clearly observed. Then mice were anaesthetized by isoflurane and the baseplate was fixed with dental cement.

Calcium imaging in freely moving mice. Mice were habituated to head-fixation and the microscope was connected to the baseplate when the animal was head-fixed followed by 30 min acclimatization before imaging sessions. Fluorescence images were acquired at 10 Hz and the LED power was set 10–35% (0.1–0.35 mW) with analogue gain 3–4. To compare the Ca^{2+} activity in different test sessions, the image acquisition parameters were set to the same values. Animal behaviour was recorded by a top mounted camera (Basler) (30 Hz). A synchronization signal between the miniature microscope and camera was recorded by a signal acquisition system (Neuralynx).

Ghrelin (1 μg per g) and saline injections (i.p.) were performed on *ad libitum* fed mice. Food-restricted mice (80–85% initial body weight) underwent chow food, false food, and Pavlovian trace conditioning tests. Chow food identical to that in the home cage (see Mice section of Methods). False food was a similar sized wood block/foam plug. For feeding experiments, either object was placed into the test arena 1.5 min after the onset of imaging sessions. For short exposure tests, the object was removed 1 min after delivery. Events of delivery, contact, leaving, and removal were manually marked from the behaviour video. During Pavlovian trace conditioning test, a 200 ms auditory (12 kHz) and a visual (blue light) compound conditioned stimulus was randomly presented 60–90 s after the onset of imaging sessions. A lickometer spout was extended 800 ms after the cue with 30 s to access the spout which delivered a palatable liquid food (Ensure). Lick events were recorded by the Neuralynx system.

Calcium image analysis. All image analyses were performed in ImageJ and Matlab. Because AGRP neurons are at the base of the brain and next to the third ventricle, the mechanical drift between GRIN lens and brain tissue included some nonlinear distortions. Movement was corrected in Janelia Computer Cluster by a custom Matlab script using an open source toolkit ANTs[46] (http://picsl.upenn.edu/software/ants/). The movement-corrected images were cropped to remove the margin values filled by the registration. This movement correction algorithm allowed image analysis even when mice were chewing chow food pellets, typically a difficult case for brain imaging (for example, Fig. 4f does not show movement artefacts associated with initiation or cessation of chewing). As an example that calcium activity reduction observed during eating is not due to movement artefacts, Fig. 4f shows a neuron (neuron 1) that increases activity and is surrounded by other neurons (3, 10, 11, 12, see Fig. 4d) that reduce activity, indicating that the response properties are due to specific neuron dynamics and not a generalized movement artefact, which would affect the entire local area of the neurons.

To extract calcium indicator fluorescence responses associated with individual neurons, a cell-sorting algorithm[47] based on principal component analysis and independent component analysis was used to automatically compute ROI spatial filters that were applied to the aligned imaging data. Separate spatial filters for each cell consisted of a weight-matrix with values between zero and one that was used to compute the fractional contribution of each pixel to the calculation of calcium fluorescence. Background fluorescence was subtracted from cropped images using ImageJ background subtraction function. Ca^{2+} activity of individual cells within ROI spatial filters were extracted from the background subtracted images. $\Delta F/F_0$ was calculated as $(F-F_0)/F_0$, where F_0 is the lowest 5% of the fluorescence signal in image sessions on one day. Normalized $\Delta F/F_0$ was used to transform the range of $\Delta F/F_0$ to [0 1] by the equation: $(\Delta F/F_0 - \min(\Delta F/F_0))/(\max(\Delta F/F_0) - \min(\Delta F/F_0))$.

ROI spatial filters identified from the same field of view were often different under different test sessions. To compare Ca^{2+} activity in such situations, only the intersection of the ROI spatial filters were used to do further analyses and comparisons. Under *ad libitum* fed conditions, Ca^{2+} activity is quite low, and only few ROIs can be detected by the cell-sorting algorithm. To compare AGRP Ca^{2+} activity between *ad libitum* fed and other conditions, the ROIs detected under other test conditions with higher GCaMP6 fluorescence were manually mapped to images from the *ad libitum* fed condition.

To test whether Ca^{2+} activity of individual neurons was changed after ghrelin injection or chow food delivery, 30% of initial mean baseline fluorescence was chosen as a threshold. Neurons with mean fluorescence changes after the experimental manipulation that were more than the threshold, either decreasing or increasing fluorescence, were categorised as changed. For ghrelin administration, the mean fluorescence (from 1 min recording) after ghrelin injection (10 min post-injection) was compared to 1 min mean pre-injection GCaMP6 fluorescence. For food delivery, the mean fluorescence (from 1 min recording) from immediately after food delivery was compared to 1 min mean pre-food GCaMP6 fluorescence.

Evoked water consumption with SFO neuron activation. For cell type-specific evoked water consumption, *Nos1* was identified as a marker for SFO neurons by inspection of the Allen Brain Atlas. SFO^{hM3Dq} and $SFO^{NOS1\text{-}ChR2}$ mice were housed with *ad libitum* food and water and were transferred into a behavioural arena (Coulbourn Instruments) with food pellets (20 mg each) delivered through an automatic pellet dispenser as previously described. Water was supplied through a ball bearing-gated metal spout that was fed by a water bottle. Licking was detected by beam breaks and calibration experiments showed lick volume was 1.35 ± 0.21 ml (n = 2 mice, mean ± s.d.).

All evoked drinking tests were performed during the early light period. Water intake was recorded for 1 h before the onset of neuron activation to establish a baseline drinking rate. This was followed by a chemogenetic or optogenetic SFO neuron stimulation period. For SFO^{hM3Dq} mice, baseline water consumption was measured for 1 h followed by clozapine-*N*-oxide (Enzo Life Sciences) injection (i.p., 2.5–5 mg per kg). Unless otherwise noted, mice had free access to food and consumption was also measured.

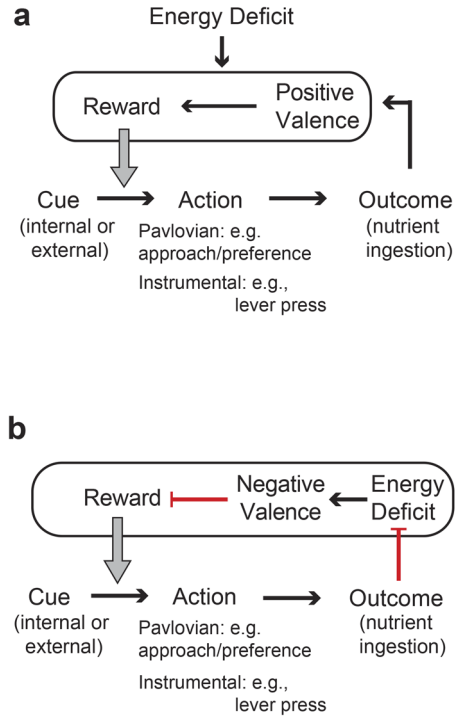
Progressive ratio 3 reinforcement schedule for water. SFO^{hM3Dq} mice that exhibited elevated water consumption in response to CNO (2.5–5 mg per kg, i.p.) were used for instrumental lever-press training (mice that did not show elevated water consumption were determined, post hoc, to lack hM3Dq-mCherry expression in the SFO). Training and tests were conducted in operant conditioning chambers (Coulbourn) with two levers, one active and one inactive, at either side of the water dispenser. A pinch valve faucet controlled opening of a syringe tube from which the water was dispensed. For each trial, after reaching lever-press criteria on the active lever, a water reward (4–6 μl) was delivered by releasing the pinch valve (NResearch) and was controlled by Graphic State Software (Coulbourn instruments). Lever pressing on the inactive lever was monitored but did not lead to water delivery.

For training, SFO^{hM3D} mice were water restricted (1 ml per day) and trained in 30 min daily FR1 sessions until they performed at least 250 lever presses in a session for 3 consecutive days. They were then trained on a progressive ratio 3 (PR3) reinforcement schedule (each water delivery reinforcer required 3 additional lever presses than the previous reinforcer) in 1 h sessions for three days. Following training, mice were rehydrated by *ad libitum* water access for 1 week. Mice were then tested on a PR3 reinforcement schedule in a 1 h session following CNO (2.5–5 mg per kg) or saline injection. Breakpoint was defined as the last ratio completed before 5 min has passed without earning a water reinforcer.

Statistics. Values are means ± s.e.m. Pairwise comparisons were calculated by unpaired or paired two-tail Students *t*-test or ANOVA. When equal variance assumptions were violated, nonparametric ANOVA on ranks test was used (see Extended Data Table 1). Post hoc multiple comparisons used Holm-Sidak correction. Statistical analyses and linear regressions were performed using SigmaPlot (Systat) or Matlab. Sample sizes were chosen to cover high and low viral transduction levels. Viral transduction efficiency for mice in electrical

activity perturbation experiments (Figs 1 and 2) was determined post hoc, which effectively blinded experimenters to the group identity (high or low transduction efficiency) for each subject. Results of statistical tests are summarized in Extended Data Table 1. NS, $P > 0.05$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. No statistical methods were used to predetermine sample size.

40. Atasoy, D., Aponte, Y., Su, H. H. & Sternson, S. M. A. FLEX switch targets channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28,** 7025–7030 (2008).
41. Tsai, H. C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324,** 1080–1084 (2009).
42. Branson, K., Robie, A. A., Bender, J., Perona, P. & Dickinson, M. H. High-throughput ethomics in large groups of *Drosophila*. *Nature Methods* **6,** 451–457 (2009).
43. Stamatakis, A. M. & Stuber, G. D. Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance. *Nature Neurosci.* **15,** 1105–1107 (2012).
44. Hagan, M. M. *et al.* Long-term orexigenic effects of AgRP-(83–132) involve mechanisms other than melanocortin receptor blockade. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **279,** R47–R52 (2000).
45. Krashes, M. J., Shah, B. P., Koda, S. & Lowell, B. B. Rapid versus delayed stimulation of feeding by the endogenously released AgRP neuron mediators GABA, NPY, and AgRP. *Cell Metab.* **18,** 588–595 (2013).
46. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12,** 26–41 (2008).
47. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63,** 747–760 (2009).



Extended Data Figure 1 | Models for homeostatic regulation of learning food preferences and food-seeking behaviours. **a**, The relationship between internal or external cues and Pavlovian approach or instrumental food-seeking actions is strengthened by nutrient ingestion. Nutrients have intrinsically positive valence[7] (rewarding), and energy deficit enhances the reward value of outcomes associated with food intake. **b**, Model of food preference and food-seeking in which learning involves reducing an energy deficit internal state that has negative valence. The relationship between internal or external cues and food preferences or food-seeking actions is strengthened by nutrient ingestion outcomes that reduce energy deficit and associated negative valence (red bar arrows are inhibitory). Conversely, the relationship between internal or external cues and food preference or food-seeking actions is weakened if outcomes do not reduce energy deficit.

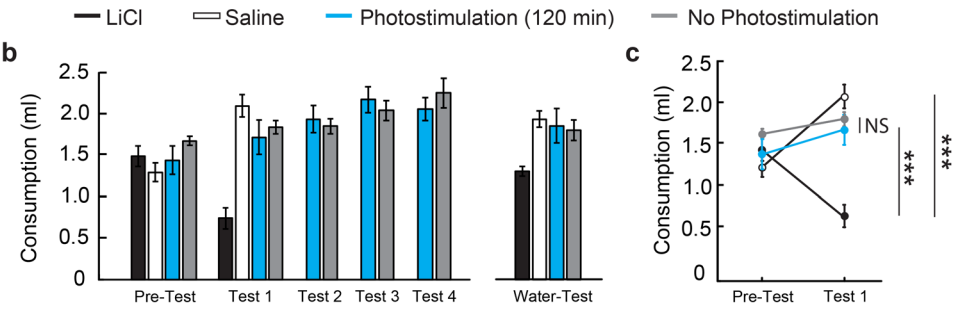
a

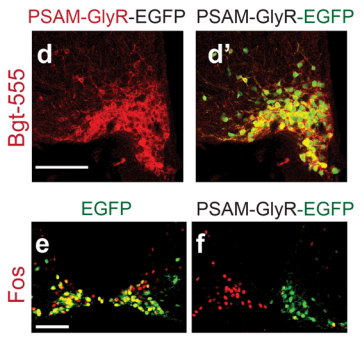
LiCl conditioned taste aversion protocol

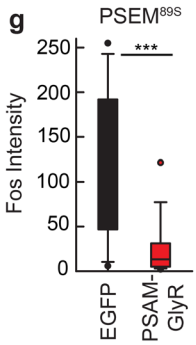
Day 1-4	Day 5	Day 6	Day 7
Habituation (water)	Conditioning (0.15% saccharin followed by LiCl or saline injection)	Test (0.15% saccharin)	Control Test (water)

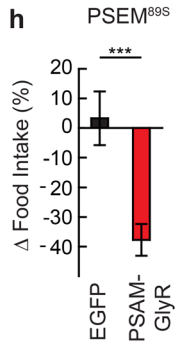
AGRP neuron activation conditioned taste aversion protocol

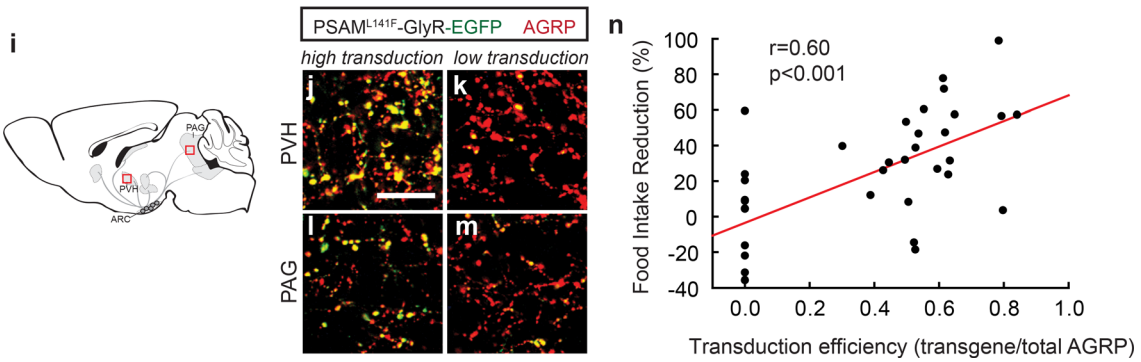
Day 1-4	Day 5-12 (Conditioning and Test x 4)		Day 13
Habituation (water)	Conditioning (0.15% saccharin followed by Photostim or No Photostim)	Test (0.15% saccharin)	Control Test (water)



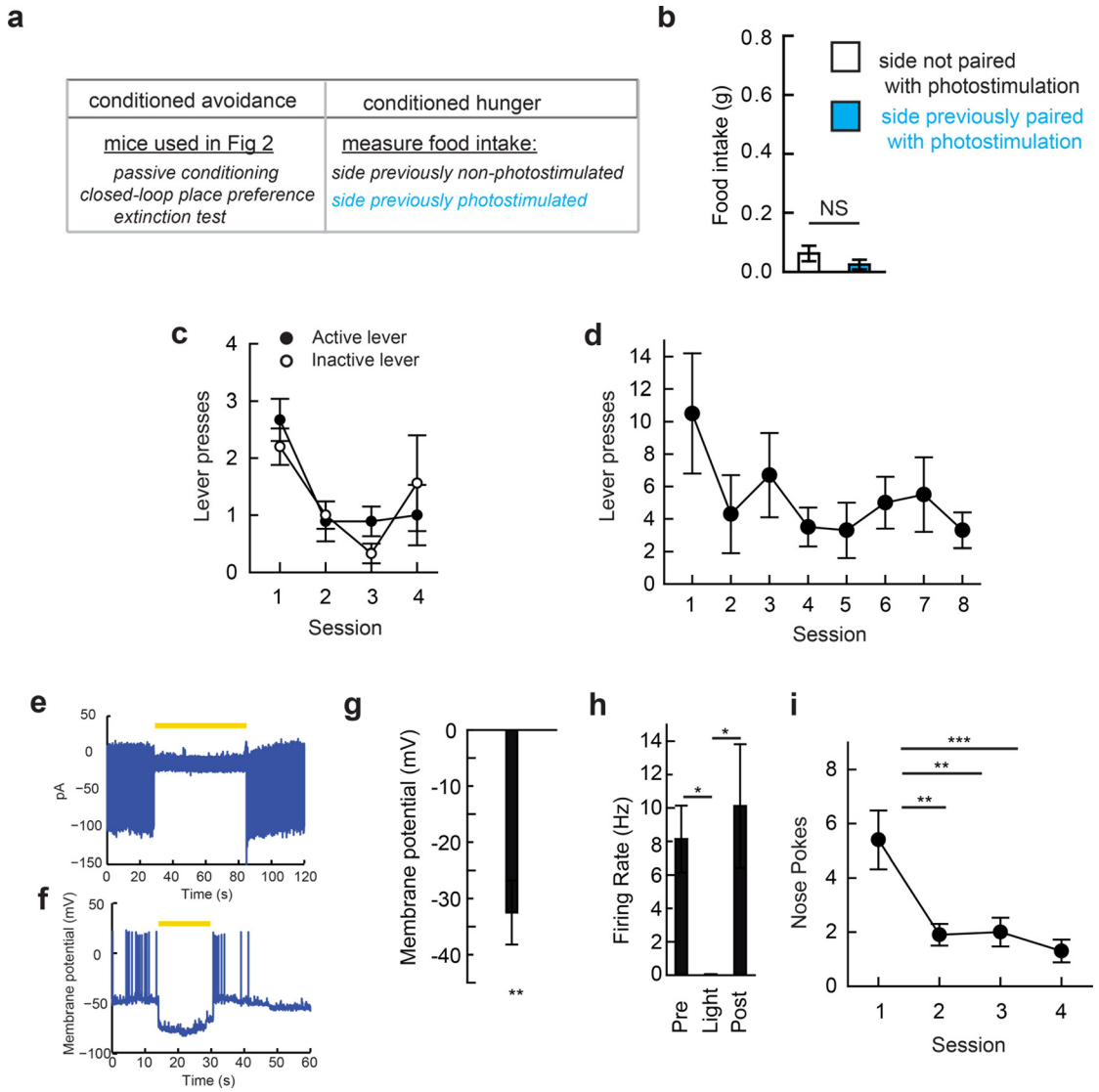




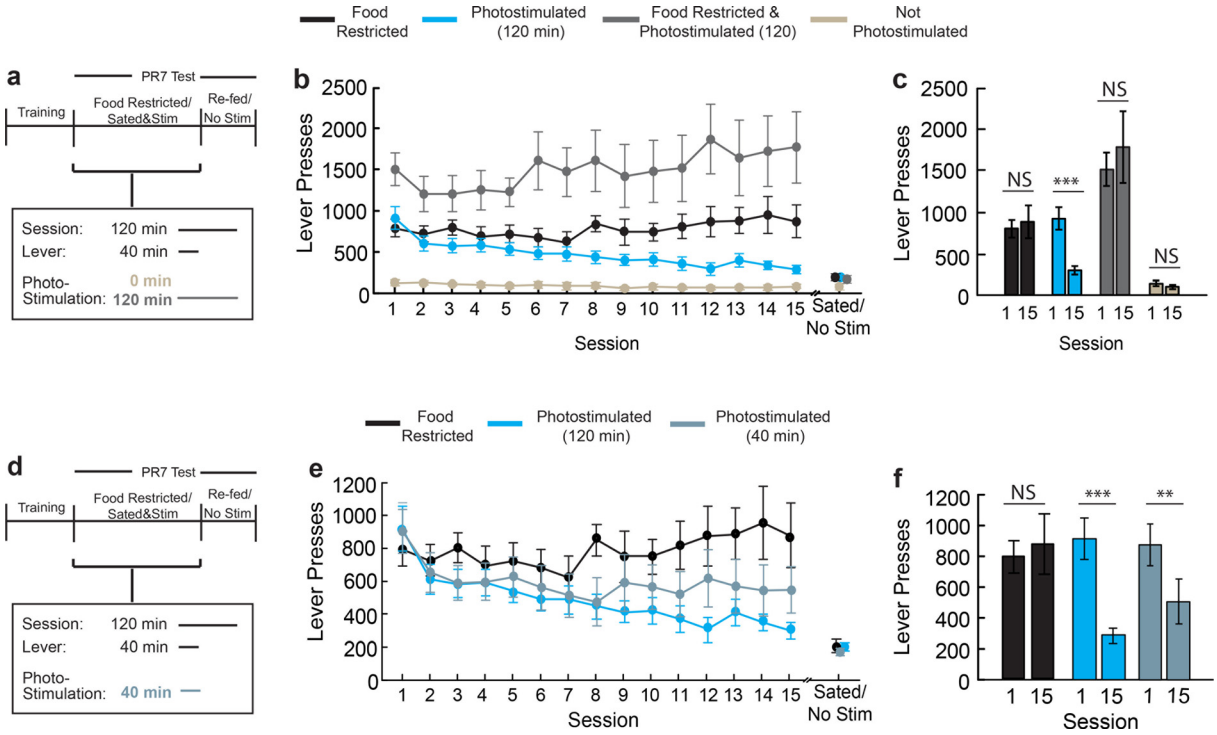




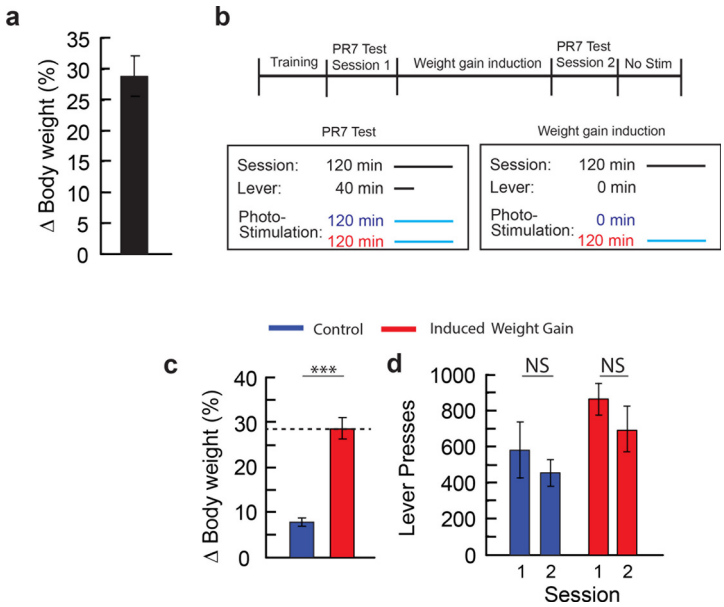
Extended Data Figure 2 | AGRP neuron activation does not condition taste aversion, and feeding reduction correlates with proportion of AGRP neurons inhibited. **a**, Experimental design for conditioned taste aversion experiments. Mice were water restricted and habituated to drink water from a spout during 20 min sessions. Four groups of mice were then allowed to consume a tastant (0.15% saccharin solution) for 20 min (pre-test) and immediately following this session, they were exposed to a conditioning agent (LiCl, saline, 120 min AGRP neuron photostimulation, or $AGRP^{ChR2}$ mice attached to an optical fibre but not phostostimulated; all $n = 6$ mice). The next day, mice were tested for consumption of the saccharin solution (test 1). For AGRP neuron photostimulated and non-photostimulated groups, conditioning and testing was extended with an additional three conditioning and test sessions. The day following the last testing session for each group, water consumption was also measured (water test). **b**, **c**, Consumption of tastant solution for all sessions (**b**) and comparison for pre-test and test 1 session (**c**). **d**, Confocal micrographs of Cre recombinase-expressing AGRP neurons transduced with rAAV-*Syn*-FLEX-$PSAM^{L141F}$-GlyR-IRES-eGFP. Alexa555-conjugated-Bungarotoxin (Bgt-555) labels $PSAM^{L141F}$-GlyR, which co-localizes with eGFP. Scale bar, 100 μm. **e**, **f**, Fos immunofluorescence in the ARC of mice treated with $PSEM^{89S}$ during the first 4 h of the dark period without access to food. $AGRP^{eGFP}$ mice (**e**) show high levels of Fos in AGRP neurons, and $AGRP^{PSAM\text{-}GlyR}$ mice (**f**) express low levels of Fos in neurons that express PSAM-GlyR (right side); non-transduced neurons (contralateral side) express high levels of Fos. Scale bar, 100 μm. **g**, Fos immunofluorescence intensity in AGRP neurons from $AGRP^{PSAM\text{-}GlyR}$ or $AGRP^{eGFP}$ mice after $PSEM^{89S}$ treatment during the first 4 h of the dark period without access to food ($n = 3$ mice per condition, $n > 50$ nuclei per condition). **h**, Change in food intake for $AGRP^{eGFP}$ mice ($n = 12$) or $AGRP^{PSAM\text{-}GlyR}$ mice ($n = 23$) treated with $PSEM^{89S}$ during the first 4 h of the dark period relative to saline injected on successive days. **i**, Diagram of AGRP neuron axon projection fields showing from where transduction efficiency was calculated. **i**–**m**, After rAAV-*hSyn*-FLEX-*rev*-$PSAM^{L141F}$-GlyR-IRES-eGFP transduction of *Agrp-IRES-Cre* mice, measurement of eGFP transduction efficiency in AGRP boutons in the PVH (**i**, **k**) and PAG (**l**, **m**). High transduction efficiency (>50% in AGRP boutons) is shown (**i**, **i**) in comparison to low transduction efficiency (<50% in AGRP boutons) (**k**, **m**). Scale bar, 20 μm. **n**, Food intake reduction for mice treated with $PSEM^{89S}$ is correlated with the transduction efficiency of rAAV-*hSyn*-FLEX-*rev*-$PSAM^{L141F}$-GlyR-IRES-eGFP in AGRP neurons (eGFP transduced boutons per total AGRP boutons) ($n = 35$ mice). NS, $P > 0.05$, $^{***}P < 0.001$. Values are means ± s.e.m.



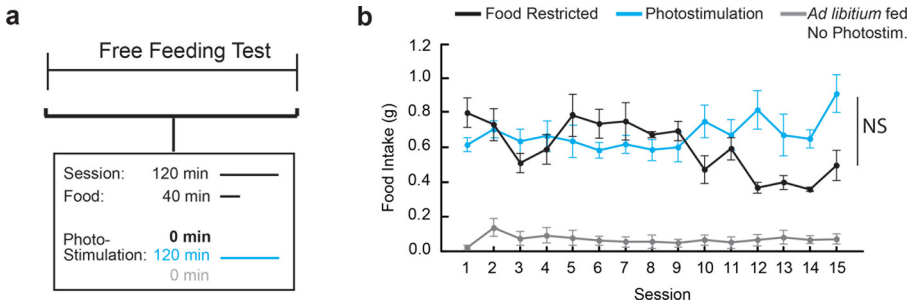
Extended Data Figure 3 | AGRP neuron activation does not condition appetite or reinforce instrumental responding. **a**, Experimental design to test conditioned appetite. After closed-loop place preference and extinction testing (Fig. 2), AGRP^{ChR2} mice showed reduced occupancy in the photostimulation-paired side of the chamber. Avoidance in extinction indicated conditioning to offset of a negative-valence signal from AGRP neurons. An alternative hypothesis is that induction of food-seeking on the photostimulation side in the absence of food led the mouse to seek food. Because photostimulation was stopped when the mouse passed to the other side of the chamber, this might increase occupancy on the non-photostimulated side. However, this is not consistent with the increased avoidance of the previously photostimulated side in extinction (Fig. 2k) unless the contextual cues previously associated with photostimulation conditioned increased appetite. To test whether conditioned avoidance might be associated with conditioned hunger, we measured food intake in *ad libitum* fed mice after closed-loop place preference and extinction tests in Fig. 2g–k on each side of the apparatus in the absence of photostimulation. **b**, Mice did not show conditioned food consumption on the previously photostimulated side (paired *t*-test, $n = 8$ mice). This indicates that avoidance observed in extinction was not a consequence of food-seeking behaviours being differentially engaged on one side of the apparatus. **c**, **d**, Cessation of AGRP neuron photostimulation did not condition instrumental responding. **c**, Nose pokes by *ad libitum* fed AGRP^{ChR2} mice ($n = 9$) during photostimulation, where a nose poke resulted in a 20 s pause in light pulses for each behavioural session. Nose pokes reduced across sessions indicating the absence of instrumental conditioning. Filled circles: active port, empty circles: inactive port. **d**, For *ad libitum* fed AGRP^{ChR2} mice previously trained to hit a lever for food, lever presses during photostimulation, where a lever press gives a 20 s pause in light pulses for each behavioural session (repeated measures ANOVA $F_{(7,40)} = 1.19$, $P = 0.330$; $n = 8$ mice). **e–h**, Optogenetic silencing with Arch (550–600 nm, 8–11 mW per mm^2). **e**, Cell-attached recording of AGRP neuron firing rate in brain slices from *Agrp-IRES-Cre;Ai35d* (AGRP^{Arch}) mice during light illumination. **f**, Whole cell recording of AGRP^{Arch} during optogenetic inhibition. **g**, Membrane potential change in AGRP neurons expressing Arch during light illumination ($n = 6$). **h**, AGRP neuron firing rate during optogenetic inhibition of Arch-expressing AGRP neurons ($n = 4$). **i**, Optogenetic silencing of AGRP neurons in food-restricted mice did not condition free operant instrumental responding. Nose pokes by AGRP^{Arch} mice resulted in 60 s of 561 nm light delivered to an optical fibre over the ARC. Nose poking reduced over multiple sessions (ANOVA $F_{(3,24)} = 7.835$, $P < 0.001$; $n = 7$ mice), indicating that silencing AGRP neurons did not directly reinforce instrumental responding. NS, $P > 0.05$ Values are means ± s.e.m.



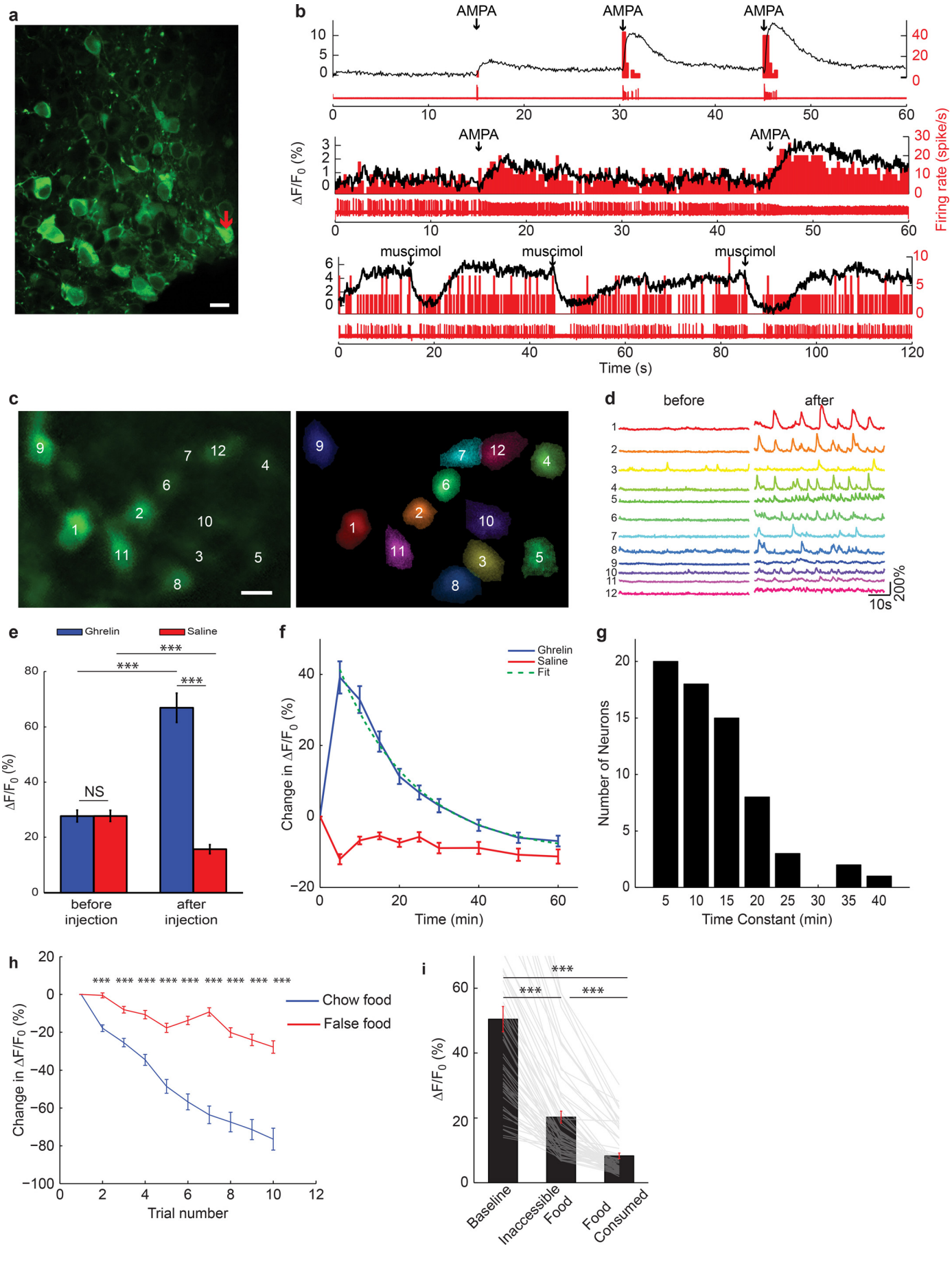
Extended Data Figure 4 | Lever pressing for food is sensitive to AGRP neuron photostimulation duration. **a**, Experimental design of progressive ratio 7 lever-press experiment from Fig. 3 for a food-restricted AGRP neuron photostimulated group and an *ad libitum* fed non-photostimulated group. The two additional groups of mice were trained to lever press in food-restriction on a PR7 reinforcement schedule. For the food-restricted with photostimulation group, mice were maintained on food-restricted and tested with PR7 reinforcement tests over 15 sessions with photostimulation. Each session was 2 h, where levers were available for the first 40 min of the session, and photostimulation was delivered for the length of the session (120 min, grey). Mice were then *ad libitum* re-fed and tested on a non-photostimulated PR7 session. For the *ad libitum* fed non-photostimulated group, mice were *ad libitum* re-fed following lever-press training and tested with PR7 reinforcement tests over 16 sessions, with no photostimulation delivered (beige). **b**, Lever presses for each PR7 session for food-restricted AGRP neuron photostimulated mice (grey, $n = 11$) mice and *ad libitum* fed non-photostimulated mice (beige, $n = 8$). For comparison, data are shown for food-restricted and 120 min photostimulated groups that are reproduced from Fig. 3b. **c**, Lever presses on first (1) and last (15) sessions in PR7 test for food-restricted with photostimulation mice (grey) mice and sated no photostimulation mice (beige). Also shown are data for food-restricted and 120 min photostimulated groups that are reproduced from Fig. 3c. **d**, Experimental design of progressive ratio 7 lever-press experiment from Fig. 3 for a 40 min photostimulation group. One additional group of mice was trained to lever press in food-restriction on a PR7 reinforcement schedule. Mice were then *ad libitum* re-fed and tested with PR7 reinforcement tests over 15 sessions. Each session was 2 h, where levers were available for the first 40 min of the session, and photostimulation was delivered only while levers were available (grey). A non-photostimulated PR7 session was also performed after the fifteenth test session. **e**, Lever presses for each PR7 session for 40 min photostimulated (grey, $n = 12$) mice. Also shown are data for food-restricted and 120 min photostimulated groups that are reproduced from Fig. 3b. **f**, Lever presses on first (1) and last (15) sessions in PR7 test for 40 min photostimulated mice (grey). Also shown are data for food-restricted and 120 min photostimulated groups that are reproduced from Fig. 3c. NS, $P > 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Values are means ± s.e.m.



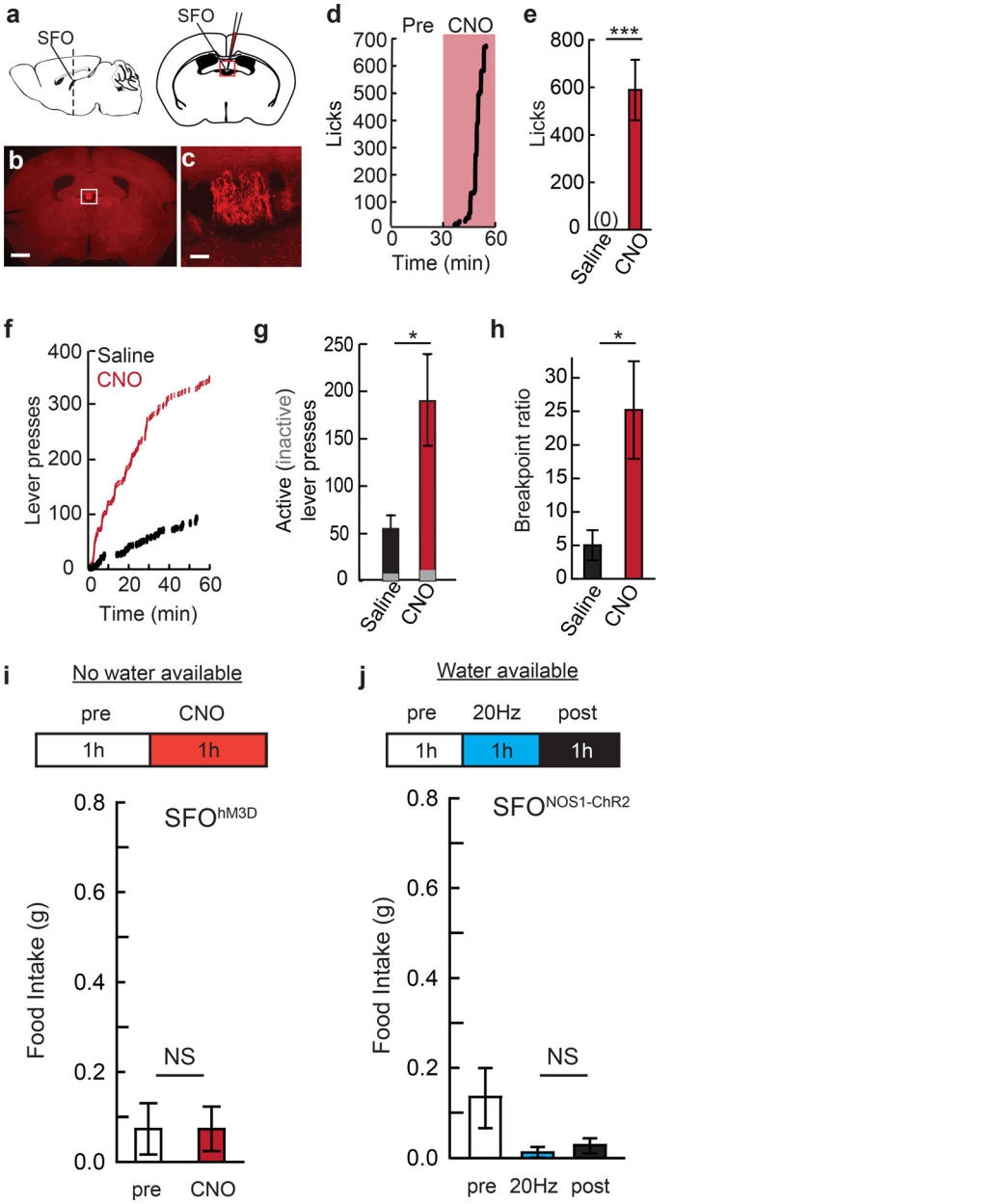
Extended Data Figure 5 | AGRP neuron-associated body weight increase does not suppress AGRP neuron-evoked food-seeking. a, Weight gain for the 120 min AGRP neuron photostimulated ($n = 11$) group in the PR7 experiment (from Fig. 3) after 15 sessions. Weight gain is due to eating after the test session when the mouse is returned to the homecage and is associated with long-lasting effects from release of AGRP[44]. Previous experiments have shown that AGRP is not responsible for the acute feeding behaviour investigated in this study[4,5,45]. However, metabolic changes associated with weight gain could be an alternative cause of reduced instrumental food-seeking shown in Fig. 3. To test the effect of weight gain in mice trained to lever press for food on a PR7 reinforcement schedule, we induced weight gain without the negative reinforcement extinction protocol from Fig. 3. **b**, Experimental design of progressive ratio 7 lever-press experiment with AGRP neuron photostimulation-induced weight gain but lacking disruption of negative reinforcement during food-seeking. $AGRP^{ChR2}$ mice were trained under food deprivation to lever press under a PR7 schedule for food pellets. After training, both groups were *ad libitum* re-fed, and the mice were divided into two groups: (1) control mice with no induction of weight gain (blue) and (2) the induced weight gain group (red). Both groups were then tested on a PR7 reinforcement schedule under AGRP neuron photostimulation conditions (PR7 test 1). Following this session, a photostimulation-induced weight gain protocol was initiated for the second group. Mice received one 2 h experimental session per day, where they were photostimulated for the whole experimental session and body weight was monitored daily. During these sessions, levers were not available, but free food was provided during these sessions (the amount of food was matched in quantity to the average amount of food acquired by the 120 min photostimulation group under the PR7 experiment from Fig. 3 for the corresponding session). The photostimulation-induced weight grain protocol was conducted for 22 consecutive days, which was required for body weight gain to be comparable to levels acquired by the 120 min AGRP neuron photostimulation group in the PR7 experiment (~28%) from Fig. 3. Control mice were tethered to a fibre but did not receive photostimulation, otherwise they received the same experimental manipulation as induced weight gain mice (access to the same amount of food), and their body weight was also monitored. After the induced weight gain group achieved a 28% weight gain, a second PR7 test was conducted for both groups in the same manner as the first one. **c**, Per cent body weight change for control (blue, $n = 6$) and induced weight gain (red, $n = 6$) mice. Grey dotted line: per cent body weight change for photostimulated mice in PR7 experiment from Fig. 3. **d**, Lever presses for control (blue) and AGRP neuron photostimulation-induced weight gain (red) mice on first (1) and second (2) PR7 test, prior and after weight gain induction protocol, respectively. There was no significant reduction in lever pressing between PR7 sessions within either group. NS, $P > 0.05$, $^{***}P < 0.001$. Values are means ± s.e.m.



Extended Data Figure 6 | Free food consumption is not reduced with repeated daily AGRP neuron photostimulation sessions. **a**, Experimental design of free feeding experiment over repeated sessions. Three groups of $AGRP^{ChR2}$ mice were tested on a 15 session free feeding protocol (no lever pressing required) either under food restriction (black), *ad libitum* fed AGRP neuron photostimulated (cyan), or *ad libitum* fed without AGRP neuron photostimulation (grey) conditions. On each day, mice received one 2 h session where food was freely available for the first 40 min of the session. AGRP neuron photostimulated group received photostimulation for the entire 2 h session (cyan). **b**, Food intake for each session of the free feeding experiment for food restricted (black, $n = 6$), AGRP neuron 120 min photostimulated (cyan, $n = 6$), and no photostimulation (grey, $n = 6$) groups. NS, $P > 0.05$. Values are means ± s.e.m.



Extended Data Figure 7 | Calcium imaging of AGRP neurons in freely moving mice. **a**, Projection of confocal images of AGRP neurons from brain slices after mice expressed GCaMP6s for 10 months after viral injection. A total of >99.5% neurons show nuclear exclusion of GCaMP6s, indicating good cell health. Red arrow, rare example of filled nucleus. Scale bar, 15 μm. **b**, Characterization of the relationship between action potential firing rate (cell attached recordings) and change of GCaMP6f fluorescence activity in brain slices by puffing AMPA for activation (top, middle) or muscimol for inhibition (bottom). **c**, Epifluorescence images of $AGRP^{GCaMP6f}$ neurons (left) from *ad libitum* mice after ghrelin injection by deep-brain calcium imaging and their ROI spatial filters (right) for image analysis. Scale bar, 15 μm. **d**, For freely moving *ad libitum* fed mice during *in vivo* imaging, fluorescence traces of individual $AGRP^{GCaMP6f}$ neurons in **c** before and after ghrelin injection (fluorescence responses separated in time by 4 min, during which time ghrelin was injected). **e**, Changes in mean Ca^{2+} activity before and 4 min after ghrelin/saline injections (90 neurons, 4 *ad libitum* fed mice). **f**, Time course of changes in mean Ca^{2+} activity after ghrelin (blue) or saline (red) injection (90 neurons, 4 mice). Green dashed line, exponential fit. **g**, Distribution of individual time constants for decline of ghrelin-mediated fluorescence increase for individual neurons showing goodness of fit >0.85 (67/90 neurons, 4 mice). **h**, Baseline GCaMP6f fluorescence at the start of each trial before 1 min exposure to food/wood in each trial. **i**, GCaMP6f fluorescence comparing initial baseline activity, exposure to an inaccessible but visible food outside the cage, and subsequent consumption of food (60 neurons, 2 mice). NS, $P > 0.05$, $^{***}P < 0.001$. Multiple comparisons with Holm's correction. Values are means ± s.e.m.



Extended Data Figure 8 | SFO neuron-evoked water seeking and consumption **a**, Schematic of injection targeting of hM3Dq-mCherry to SFO neurons. **b**, Epifluorescence image of mCherry fluorescence in a coronal section containing the SFO (box) targeted stereotaxically by co-injection of rAAV-*hSyn*-Cre and rAAV-*Ef1a*-DIO-hM3Dq-mCherry. Scale bar, 1 mm. **c**, Confocal micrograph of SFO neurons co-transduced with rAAV-*hSyn*-Cre and rAAV-*Ef1a*-DIO-hM3Dq-mCherry. Scale bar, 100 μm. **d**, Number of licks for a representative SFO^{hM3Dq} mouse during evoked water consumption following activation of SFO neurons by CNO injection. **e**, Number of licks for SFO^{hM3Dq} mice following saline or CNO injection ($n = 5$ mice). **f**, Cumulative lever pressing for a SFO^{hM3Dq} mouse following injection of CNO (red) or saline (black). **g**, **h**, For SFO^{hM3Dq} mice, lever presses (red/black: active lever, grey: inactive lever) (**g**) and breakpoint reinforcement ratio (**h**) on a PR-3 water reinforcement schedule following either saline or CNO injection ($n = 5$ mice). **i**, Top, experimental design to test if activation of SFO neurons can elicit food consumption in the absence of water. SFO^{hM3D} mice were presented with access to food but not water for one hour (pre), which was followed by CNO injection, and food intake was measured for an additional hour. Bottom, food intake by SFO^{hM3D} mice that lack access to water before (pre) and after the application of CNO (paired *t*-test, $n = 3$). **j**, Top, experimental design to test if activation or offset of SFO neurons elevate food consumption behaviour. $SFO^{NOS1\text{-}ChR2}$ mice had access to food and water, and both were measured before (1 h, pre), during (20 Hz, 1 h), and after photostimulation (1 h, post). Bottom, food consumed by $SFO^{NOS1\text{-}ChR2}$ mice before (pre), during (20 Hz), or after (post) photostimulation (paired *t*-test, $n = 5$). NS, $P > 0.05$. Values are means ± s.e.m.

Extended Data Table 1 | Results of statistical analysis

Figure	Sample size (n)	Statistical Test	Values
1f	EGFP: 6; ChR2:8	Unpaired t-test	p=0.03
1j	low: 13; high: 16	Unpaired t-test	p<0.001
1k	low: 13; high: 16	Unpaired t-test	p=0.005
1l	29	Pearson Correlation	r=0.74, p<0.001
1m	<30%: 15 >30%: 14	2-WAY RM ANOVA *factor 1: (group) <30% v. >30%* *factor 2: pre/post session* *Interaction: group x session* Post hoc multiple comparisons, with Holm-Sidak corrections	F(1,27)=5.6, p=0.025 F(1,27)=3.9, p=0.06 F(1,27)=28.5, p<0.001 <30%: p=0.041; >30% p<0.001 pre: p=0.36; post: p<0.001
2c	EGFP-FR: 13; PSAM-GlyR-FR: 20 PSAMGlyR-FR: 20; AL: 9	Unpaired t-test Unpaired t-test	p=0.036 p=0.024
2d	26	Pearson Correlation	r=0.58, p=0.002
2e	35	Pearson Correlation	r=0.56; p<0.001
2f	8	Pearson Correlation	r=0.81, p=0.014
2i	12 per group	2-WAY RM ANOVA; Factor 1: (group) AGRP-ChR2 v. AGRP-GFP *Factor 2: session* *Interaction: group x session* Post hoc mult. comparisons, Holm-Sidak corrections, *Session 7*	F(1,154)=3.0, p=0.097 F(7,154)=2.3, p=0.029 F(7,154)=3.3, p=0.003 p=0.003
2j	12 per group	2-WAY RM ANOVA; Factor 1: (group) AGRP-ChR2 v. AGRP-GFP *Factor 2: session* *Interaction: group x session* Post hoc mult. comparisons. Holm-Sidak corrections. *Session 4* *Session 5; Session 6; Session 7*	F(1,154)=6.4, p=0.019 F(7,154)=3.2, p=0.004 F(7,154)=3.8, p<0.001 p=0.025 p=0.009; p=0.035; p<0.001
2k	ChR2: 12, EGFP: 12	Unpaired t-test	p=0.025
3b	Food Restriction: 11 Photostimulation: 11	2-WAY RM ANOVA; Factor 1: (Group) Restricted v. Photostim. *Factor 2: Session* *Interaction: group x session* Post hoc mult. comparisons. Holm-Sidak corrections. *Session 3* *Session 8; Session 10* *Session 11; Session 12* *Session 13; Session 14; Session 15*	F(1,280)=7.90, p=0.011 F(14,280)=1.76, p=0.045 F(14,280)=2.97, p<0.001 p=0.047 p=0.015; p=0.048 p=0.002; p<0.001 p=0.014; p=0.003; p<0.001
3c	Food Restriction: 11; Photostimulation: 11	Paired t-test	Food Restricted: p=0.682; Photostim: p<0.001
3d	Food Restriction: 11; Photostimulation: 11	Paired t-test	Food Restricted: p=0.821; Photostim: p<0.001
3e	Food Restriction: 11; Photostimulation: 11	Paired t-test	Food Restricted: p=0.385; Photostim: p=0.002
3g	Food Restriction: 11 Photostimulation: 11	2-WAY RM ANOVA; Food Restricted group; Factor 1: Session *Factor 2: Time Block* *Interaction: Session vs Time Block* Photostimulated group; Factor 1: Session *Factor 2: Time Block* *Interaction: Session vs Time Block* Post hoc multiple comparisons with Holm-Sidak corrections *Session 1: First 10 min vs Rest of Session* *Session 15: First 10 min vs Rest of Session* *First 10 min: Session 1 vs Session 15* *Rest of Session: Session 1 vs Session 15*	F(1,10)=1.29, p=0.282 F(1,10)=3.99, p=0.074 F(1,10)=5.20, p=0.046 F(1,10)=34.15, p<0.001 F(1,10)=27.94, p<0.001 F(1,10)=7.55, p=0.021 Food Restricted: p=0.357;Photostim: p=0.076 Food Restricted: p=0.009; Photostim: p<0.001 Food Restricted: p=0.045; Photostim: p=0.006 Food Restricted: p=0.824; Photostim: p<0.001
4e	61 neurons, 4 mice	Paired t-test	p<0.001
4i	110 neurons, 4 mice	RM ANOVA on RANKS Post hoc multiple comparisons (Holm-Sidak) *1st trial base v. 1st trial food* *1st trial base v. satiety* *1st trial food v. satiety*	p<0.001 p<0.001 p<0.001 p<0.001
4l	Before: 60 neurons; After: 65 neurons; 3 mice	*Unpaired t-test*	p<0.001
5d	8	Paired t-test	5Hz: p=0.002; 10Hz: p<0.001;20Hz: p<0.001
5e	Control: 6 ChR2: 12	2-WAY RM ANOVA; Factor 1: (group) NOS1-ChR2 v. C57B6/J *Factor 2: session* *Interaction: group x session* Post hoc multiple comparisons (Holm-Sidak); Session 2 *Session3;Session 4* *Session 5-7*	F(1,112)=26.2, p<0.001 F(7,112)=0.74, p=0.64 F(7,112)=4.25, p<0.001 p=0.003 p<0.001; p=0.001 p<0.001
5g	Control: 6, ChR2: 12	Unpaired t-test	p=0.003
ED 2c	LiCl:6 Saline: 6 Photostimulation: 6 No Photostimulation: 6	2-WAY RM ANOVA; Factor 1: group (Licl, Sal., Photo, No Photo) *Factor 2: session* *Interaction: group x session* Post hoc mult. comparisons (Holm-Sidak); LiCl v. Photostimulation *LiCl v. Saline* *Photostimulation v. No photostimulation*	F(3,20)=5.76, p=0.005 F(1,20)=4.47, p=0.047 F(3,20)=29.247, p<0.001 p<0.001 p<0.001 p=0.517
ED 2g	3 mice, >50 nuclei/ condition	Mann-Whitney U-Test	p<0.001
ED 2h	EGFP: 12; PSAM-GlyR: 23	Unpaired t-test	p<0.001
ED 2n	35	Pearson Correlation	r=0.60, p<0.001
ED 3b	8 mice	Paired t-test	p=0.44
ED 3c	9 mice	2-WAY RM ANOVA; Factor 1: (group) Active v. Inactive lever Factor 2: session Interaction: group x session	F(1,24)=0.571, p=0.471 F(3,24)=4.681, p=0.01 F(3,24)=1.69, p=0.196
ED 3d	8 mice	1-Way RM ANOVA	F(7,40)=1.19, p=0.330
ED 3g	6 cells	t-test	p=0.002
ED 3h	4 cells	Unpaired t-test	pre v. light: p=0.029; light v. post: p=0.029
ED 3i	7 mice	1-Way RM ANOVA	F(3,24)=7.835, p<0.001
ED 4b	Food Restriction: 11 Photostim. (120 min): 11 Food Rest & Photostim: 11; Sated No Photostim: 8	2-WAY RM ANOVA; Factor1: Group (Restr., Photo R&Photo, NoPhoto) *Factor 2: Session* *Interaction group x session*	F(3,518)=38.86, p<0.001 F(14,518)=2.719, p<0.001 F(42,518)=1.598, p=0.012
ED 4c	Food Restriction: 11 Photostim. (120 min): 11 Food Rest & Photostim: 11 Sated No Photostim: 8	Paired t-tests	Food Restricted: p=0.682 Photostim. (120 min): p<0.001 Food Rest & Photostim: p=0.44 Sated No Photostim: p=0.225
ED 4e	Food Restriction: 11 Photostim. (120 min): 11 Photostim. (40 min): 12	2-WAY RM ANOVA; Factor1: Group (Restr., Photo-40, Photo-120) *Factor 2: Session* *Interaction group x session*	F(2,434)=3.04, p=0.063 F(14,434)=3.52, p<0.001 F(28,434)=2.08, p=0.001
ED 4f	Food Restriction: 11 Photostim. (120 min): 11 Photostim. (40 min): 12	Paired t-tests	Food Restricted: p=0.682 Photostim. (120 min): p<0.001 Photostim. (40 min): p=0.002
ED 5c	6 per group	Unpaired t-test	p<0.001
ED 5d	6 per group	Paired t-test	Control: p=0.40;Induced Weight Gain: p=0.084
ED 6b	6 per group	2-WAY RM ANOVA; Factor 1: (group) Rest. v. Photostim Factor 2: session Interaction: group x session	F(1,140)=1.069, p=0.326 F(14,140)=2.380, p=0.005 F(14,140)=6.628, p<0.001
ED 7e	90 neurons, 4 mice	2-WAY RM ANOVA; Factor 1: time (before,after) *Factor 2: group (saline,ghrelin)* *Interaction: timexgroup* Post hoc mult. comparisons (Holm-Sidak):*Before: saline v ghrelin* *After: saline v ghrelin* *Saline: before v after* *Ghrelin: before v after*	F(1,88)=43.641,p<0.001 F(1,88)=0.245,p=0.622 F(1,88)=155.805,p<0.001 p=0.894 p<0.001 p<0.001 p<0.001
ED 7h	57 neurons, 2 mice	Multiple comparisons (Holm-Sidak correction); Trials 2-10	p<0.001
ED 7i	60 neurons, 2 mice	RM ANOVA on RANKS Post hoc multiple comparisons (Holm-Sidak); base v. inaccessible *base v. food; inaccessible v. food*	p<0.001 p<0.001 p<0.001; p<0.001
ED 8e	5	Paired t-test	p<0.001
ED 8g	5	Paired t-test	p=0.027
ED 8h	5	Paired t-test	p=0.028
ED 8i	3 mice	Paired t-test	p=0.95
ED 8j	5 mice	Paired t-test	p=0.48

ARTICLE

doi:10.1038/nature14446

Neural dynamics for landmark orientation and angular path integration

Johannes D. Seelig[1] & Vivek Jayaraman[1]

Many animals navigate using a combination of visual landmarks and path integration. In mammalian brains, head direction cells integrate these two streams of information by representing an animal's heading relative to landmarks, yet maintaining their directional tuning in darkness based on self-motion cues. Here we use two-photon calcium imaging in head-fixed *Drosophila melanogaster* walking on a ball in a virtual reality arena to demonstrate that landmark-based orientation and angular path integration are combined in the population responses of neurons whose dendrites tile the ellipsoid body, a toroidal structure in the centre of the fly brain. The neural population encodes the fly's azimuth relative to its environment, tracking visual landmarks when available and relying on self-motion cues in darkness. When both visual and self-motion cues are absent, a representation of the animal's orientation is maintained in this network through persistent activity, a potential substrate for short-term memory. Several features of the population dynamics of these neurons and their circular anatomical arrangement are suggestive of ring attractors, network structures that have been proposed to support the function of navigational brain circuits.

Visual landmarks can provide animals with a reliable indicator of their whereabouts[1]. In the absence of such cues, many animals track their position relative to a reference point by continuously monitoring their own motion, a process called path integration[2]. Estimates of position based purely on self-motion cues, however, can accumulate error over time. Successful navigation then, requires animals to flexibly combine these distinct sources of information[1]. In mammalian brains this process of integration is evident in head direction cells[3], which are neurons sensitive to an animal's heading relative to visual cues in its surroundings that maintain their representation of heading in total darkness using self-motion cues[4]. With their smaller brains and identifiable neurons, insects offer tractable systems to examine the integrative neural computations underlying navigation[5]. Indeed, many insects (for example, desert ants and honeybees[6,7]) are known to navigate using landmarks and path integration[1]. Experiments in a variety of insects indicate the involvement of the central complex (CX)—a brain region conserved across insects—in such behaviour. In the fruitfly, behavioural genetics experiments have suggested that the CX is required for several components of navigation, including memory for visual landmarks[8], patterns[9] and places[10], and directional motor control[11]. Electrophysiological recordings in immobilized locusts[12] and butterflies[13] have revealed a map-like representation for the orientation of electric field vectors of polarized light, which may enable sun-compass navigation[14]. Extracellular recordings from CX neurons in tethered walking cockroaches have shown encodings of turning direction[15] and of wide-field optic flow[16], a potential cue for self-motion. However, previous studies of visual responses in the CX were conducted under conditions in which insects passively viewed visual stimuli. We sought to uncover integrative neural processes relevant to navigation in the CX by allowing a tethered fly to control and respond to visual stimuli[17] while simultaneously recording its neural activity and behaviour.

We used two-photon imaging with the genetically encoded calcium indicator GCaMP6f (ref. 18) to monitor neural responses in the CX while a head-fixed fly walked on an air-supported ball within a light-emitting-diode (LED) arena[19,20] (Fig. 1a, b; see Methods). In previous experiments, we identified a subset of neurons with projections to the CX, and specifically to rings of the ellipsoid body (EB), that show strong tuning to localized visual features[20] including vertical stripes, a class of stimuli that also induce innate fixational responses in flies[21,22]. To probe how such visual information might be used within the CX we now focused on a class of columnar neurons of the CX, each of which sends dendrites to a specific wedge of the EB (Fig. 1c, d). These neurons are abbreviated here as EBw.s neurons (see Methods)[13,23–27]. We monitored the dendritic responses of the entire EBw.s population in the EB (Fig. 1e) during walking, both under closed-loop virtual reality conditions in which the rotation of visual patterns was driven by the fly's turning movement on the ball and in darkness (Extended Data Fig. 1).

Compass-like representation of landmark orientation

When flies were exposed to a single vertical stripe stimulus (Extended Data Fig. 1a), we observed a sector of activity, or bump, in the EB that rotated concurrently with the stripe as the fly turned on the ball (Fig. 1f, j, k; Supplementary Video 1, Extended Data Fig. 2a–c). The spatial extent of the visual arena (270°) was mapped to the full angular extent (360°) of the EB (Extended Data Fig. 2d, see Methods). A population vector average (PVA) of EBw.s activity (Fig. 1g, h) sufficed to reliably decode the stripe's azimuthal position relative to the fly, or, equivalently, the fly's virtual orientation relative to the stripe (Fig. 1g–i, l, see Methods), with a fly-specific angular offset (Fig. 1m). Offsets occasionally changed between trials (for example, Fly 2 and Fly 10 in Fig. 1m and Extended Data Fig. 2e), but seldom within a trial (Extended Data Fig. 2f). Such differences in the locking of the EBw.s activity bump to visual cue position suggest that EBw.s population activity cannot be a static retinotopic representation of the animal's surroundings[20].

The single-stripe EBw.s responses (Fig. 1) could result from a tuning to visual features[20], or from a more abstracted representation of the fly's orientation with respect to its environment. To distinguish between these possibilities, we asked how EBw.s population activity changes in a more complex visual scene with multiple features (Fig. 2a; Extended Data Fig. 1b). In this environment, a visual feature map would produce an EBw.s activity pattern of increased width and

[1]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147, USA.

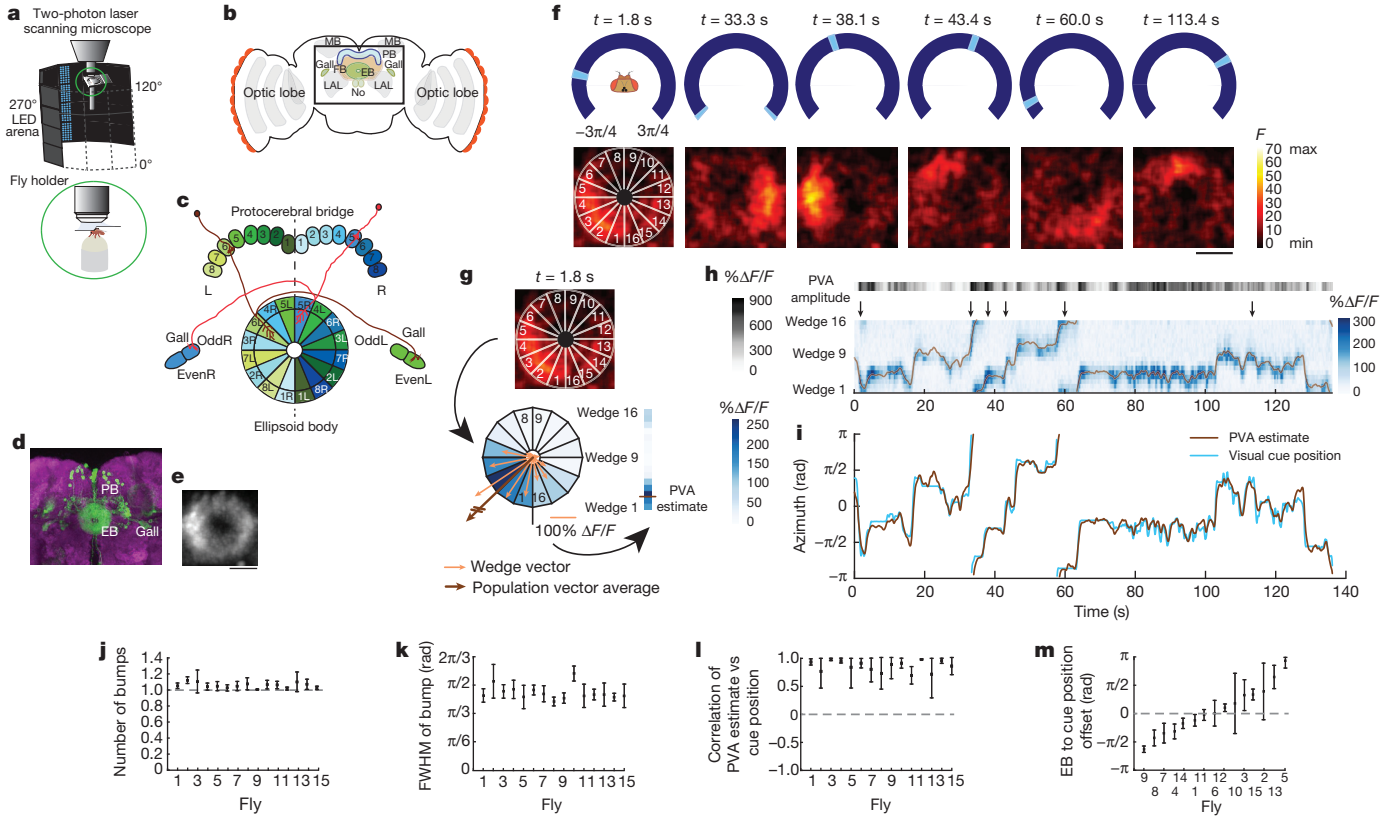


Figure 1 | Ellipsoid body activity tracks azimuth of visual cue. **a**, Schematic of setup. Inset, close-up of fly on air-supported ball (modified from ref. 20). **b**, Schematic of fly central brain and CX: ellipsoid body (EB), fan-shaped body (FB), protocerebral bridge (PB), paired noduli (NO), lateral accessory lobe (LAL) and gall (Gall). MB: mushroom body. **c**, Each EBw.s neuron receives inputs from an EB wedge and sends outputs to a corresponding PB column, and to the gall[24,26]. The PB has 18 columns[24], but EBw.s neurons only innervate the central 16. OddR, EvenR, OddL and EvenL, odd and even PB columns, right and left side of brain, respectively. **d**, GFP-labelled EBw.s neurons in a brain counterstained with nc82 (maximum intensity projection (MIP), reproduced with permission from Janelia FlyLight Image Database[23]). **e**, MIP of two-photon imaging stack (5 frames, 5 μm apart, see Methods) showing EB processes of GCaMP6f-labelled EBw.s neurons. **f**, Top, closed-loop walking with a vertical stripe. Bottom, EBw.s activity is measured in 16 regions of interest (ROIs). Sample frames from calcium imaging time series (Fly 15) showing MIP of EB activity bump (see Methods) as fly rotates visual cue around arena (top). *F*, fluorescence intensity (arbitrary units). Arrows near top of **h** indicate frame times. **g**, Steps to compute PVA based on EBw.s population activity. EB is unwrapped from Wedge 1 to Wedge 16 to display population time series in **h**. Superimposed is PVA estimate that incorporates trial-specific offset (**m**; see Methods). **h**, EBw.s fluorescence transients during a single trial (same trial as **f**). Colour scale at right. Superimposed brown line indicates PVA estimate of angular orientation of visual cue. Top, horizontal greyscale stripe shows PVA amplitude; intensity scale at left. **i**, PVA estimate of angular orientation plotted against actual orientation of visual cue (see Methods). **j**, Mean and standard deviation (s.d.) of number of activity bumps in EBw.s population activity across trials for each of 15 flies (see Methods). **k**, Mean and s.d. of full width at half maximum (FWHM) of activity bump across trials and flies (see Methods). **l**, Mean and s.d. of correlation between PVA estimate and actual orientation (pattern position) (see Methods). **m**, Mean and s.d. of angular offsets between PVA position and pattern position (see Methods, Extended Data Fig. 2e, f). All scale bars, 20 μm.

complexity. Instead, consistent with EBw.s activity representing the fly's orientation, we observed a single bump of similar width (Fig. 2a–c, Supplementary Video 2, Extended Data Fig. 3a–c), the spatial extent of the arena was once again mapped onto the EB (Extended Data Fig. 3d), and the PVA estimate of the fly's azimuth remained accurate (Fig. 2d–g, Extended Data Fig. 3e, f).

All the cues used thus far provided the fly with landmarks that uniquely define its orientation in the environment. An activity bump could thus represent either the fly's angular position within the visual scene or its orientation relative to a specific landmark within it. We next placed the fly in a visual scene with two identical vertical stripes positioned to map exactly opposite each other on the EB ring (Extended Data Fig. 1c, Extended Data Fig. 4a). Consistent with EBw.s activity representing orientation by flexibly locking to a single landmark, the resulting EBw.s representation again involved a single bump (Extended Data Fig. 4b–f; Supplementary Video 3), with the same mapping of the visual scene onto the EB as before (Extended Data Fig. 4g). PVA estimates were well correlated with the orientation of the fly in the scene (Extended Data Fig. 4h–m), regardless of which stripe was directly in front of the fly (Supplementary Video 3). In a few cases, however, we observed that EBw.s activity transitioned from one offset to another relative to the visual cues (Extended Data Fig. 5a–c, Supplementary Video 4), potentially reflecting the ambiguity inherent in determining landmark-guided orientation in an environment with multiple indistinguishable visual landmarks. Taken together, these data are consistent with a function for the EBw.s population as an internal compass that adaptively represents the fly's orientation relative to visual landmarks.

Visual landmarks dominate over self-motion cues

Our closed-loop virtual reality experiments directly couple the fly's turning movements to the rotation of the visual scene. To disambiguate the contributions of visual landmark position and self-motion cues to the EBw.s representation of the fly's orientation, we performed two sets of manipulations with a single stripe pattern. First, having observed hints of landmark-tethered activity in bump transitions in the two-stripe environment (Extended Data Fig. 5), we instantaneously shifted cue position during a period of closed-loop walking

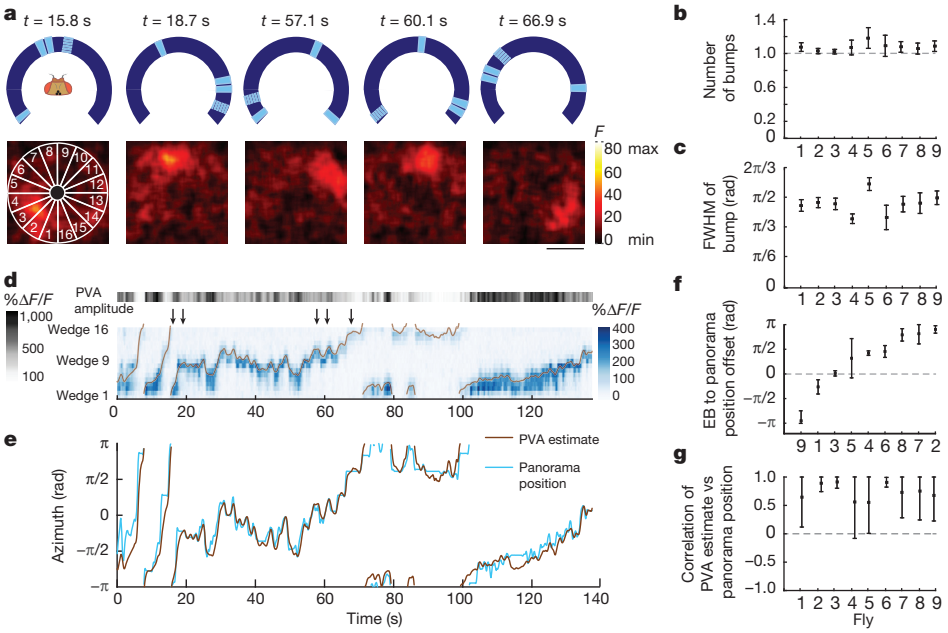


Figure 2 | Ellipsoid body is not a retinotopic map of visual scene, but represents the fly's orientation relative to visual landmarks. **a**, Closed-loop experiment in visual environment with multiple features (Fly 1 in **b**). **b**, Mean and s.d. of number of bumps across trials for each of 9 flies. **c**, Mean and s.d. of FWHM of bump. Distribution of bump widths is not significantly different from that for single-stripe stimulus (Fig. 1k); $P = 0.14$ (see Methods), mean width $= 84.9^\circ \pm 12.6^\circ$ for multiple feature trials versus $82.3^\circ \pm 11.5^\circ$ for single stripe trials. **d**, EBw.s fluorescence transients (same trial as **a**). **e**, PVA estimate of the fly's angular orientation compared to actual orientation. **f**, Mean and s.d. of angular offsets (see Methods, Extended Data Fig. 3e, f). **g**, Correlation between PVA estimate and actual orientation. Scale bar, 20 μm.

(Fig. 3a; also see Methods). If EBw.s responses arise from self-motion cues rather than landmark orientation, the abrupt shift in landmark azimuthal position should not by itself induce matching shifts in EBw.s activity. Instead, we observed that EBw.s. population activity moved to match the cue shift, preserving the initial offset between the EBw.s bump and visual cue azimuth (Extended Data Fig. 6a, b, Fig. 3b; Supplementary Video 5). Thus, local landmarks rather than self-motion cues appear to determine the position of the EBw.s bump. Interestingly, the bump did not always move instantaneously. In response to the first landmark shift in Extended Data Fig. 6b (see also Supplementary Video 5), for example, the bump rapidly tracks the shifted visual landmark, but the second abrupt displacement of the landmark elicits a much slower response.

As a second manipulation, we varied the closed-loop gain that coupled the fly's rotational movements on the ball to the movement of the visual cue (Fig. 3c). If EBw.s activity is determined by the fly's orientation relative to the landmark, the bump should move in lockstep with landmark rotation rather than with the fly's turning movements. Indeed, in almost all cases, activity in the EBw.s population faithfully followed the visual cue (Fig. 3d–g, Extended Data Fig. 6c). Consistent with this, the relationship between walking rotation and EBw.s bump movement was strongly dependent on closed-loop gain (Fig. 3h), whereas the relationship between visual cue movement and rotation of the EBw.s bump scaled only slightly with gain (Fig. 3i). Nevertheless, we occasionally observed examples of EBw.s activity being more influenced by the animal's rotation than cue movement, particularly in situations of low gain (Extended Data Fig. 6d–f). Overall, as has been observed in the mammalian head direction system[4], the EBw.s compass predominantly relies on visual landmarks, but its computation of the fly's orientation is also influenced by the animal's angular movements.

Angular path integration with no visual cues

The influence of self-motion cues on EBw.s population activity (Extended Data Fig. 6d–f) suggests that visual information can be overridden in determining compass heading. We next sought to uncover the contribution of self-motion cues in an environment that provided no visual information. A key feature of mammalian head direction cells is their ability to retain their compass-like function in the absence of visual information using path integration[4]. We searched for evidence of angular path integration, that is, tracking of orientation by self-motion cues, by imaging EBw.s population responses of flies walking in complete darkness without prior exposure to a closed-loop visual scene on the ball (Fig. 4a). As in all visually stimulated conditions, EBw.s activity in the dark settled into a single bump, and then tracked the fly's turning movements on the ball (Fig. 4b–d; Supplementary Videos 6–8, Extended Data Fig. 7a–e, h). However, the PVA estimate of orientation based on the activity in the EBw.s population often degraded over time (Fig. 4d, Extended Data Fig. 7b, f, h, i), with EBw.s activity not tracking very small or slow angular movements (Extended Data Fig. 7g, j, and Extended Data Fig. 8). Although the fly's potentially impaired ability to track its rotation when tethered on a ball may contribute to the measured drift, it likely also reflects a common limitation of path integrators in the absence of corrective feedback[4,28]. We also noted some fly-to-fly variability in the effective (measured) gain between ball rotation and EBw.s bump movement in these experiments (Extended Data Fig. 7f, i). The observation that the EBw.s system can operate at different gains raises the possibility that the compass can not only adjust to various visual closed-loop gains (as seen in Fig. 3), but also perhaps tune the gain between ball rotation and EBw.s activity through experience. However, prior exposure to specific closed-loop gains in visual surroundings only had a negligible influence on the effective gain between ball rotation and EBw.s activity in darkness (Extended Data Fig. 9). Overall, these results show that the EBw.s population performs angular path integration in darkness by relying exclusively on self-motion cues, albeit with a gradual accumulation of error in its orientation estimate.

Orientation represented by persistent activity

Having established that both visual landmark information and self-motion cues contribute to EBw.s population activity, we next asked how the system responds to the absence of both sources of orientation information. Specifically, we examined EBw.s activity during epochs when the fly stopped walking while in the dark. In almost all such cases, the EBw.s population maintained a representation of the fly's orientation (Fig. 4e–i; Extended Data Fig. 10a–f; Supplementary Videos 7 and 8). This representation persisted even when EBw.s calcium activity was low, as evident in the fact that renewed bouts of movement caused a bump to reappear in exactly the wedges expected based on the last orientation of the fly (Fig. 4h and Extended Data Fig. 10a, c, e, Supplementary Videos 7 and 8). This representation of orientation sometimes persisted for more than 30 s (Fig. 4i, Extended Data Fig. 10b, d, f). Such persistence was also a feature of EBw.s dynamics

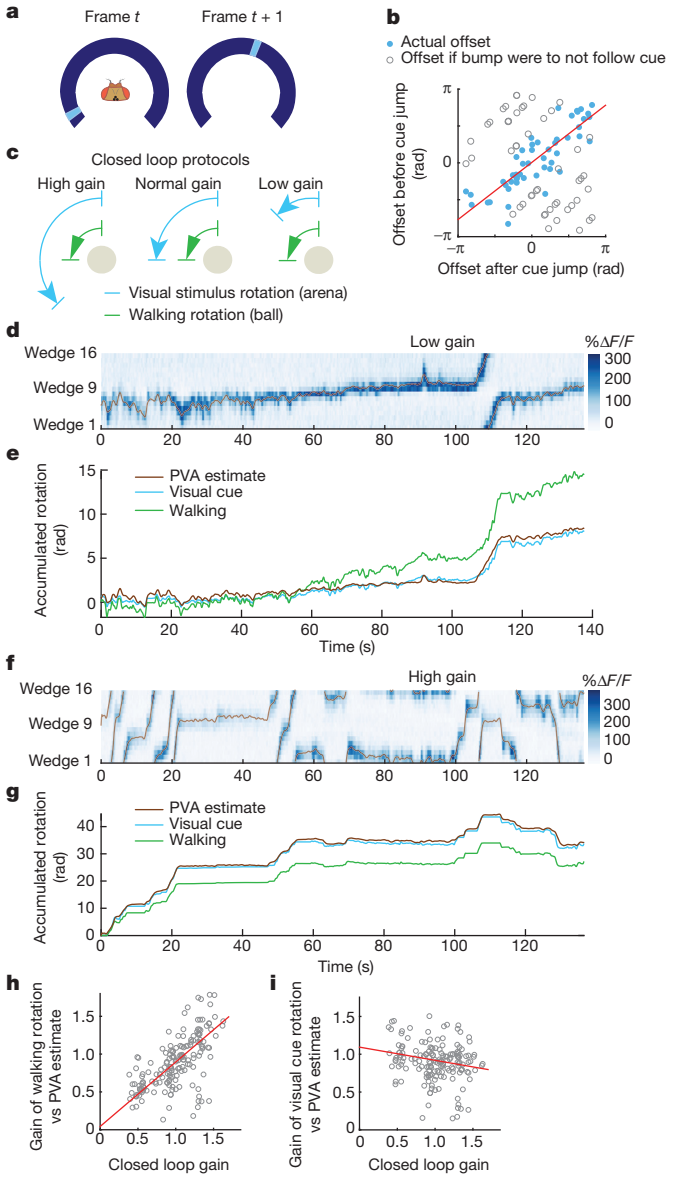


Figure 3 | EBw.s activity tracks landmark orientation cues over angular rotation when these cues are in conflict. **a**, In cue-shift experiments, fly is in closed-loop control of stripe position until cue is abruptly shifted to new position (see Methods). **b**, Offset between PVA estimate and actual orientation relative to visual cue before and after cue shift. Plot compares actual offsets with those expected if EBw.s activity did not follow cue position (see Methods). $N = 50$ shifts ($n = 6$ flies), $r = 0.85$, $p_r = 0$. Fit: slope $= 0.78 \pm 0.07$, $p_{slope} = 0$, $R^2 = 0.72$. **c**, In closed-loop gain-change experiments, ball rotation drives movement of visual stimulus with different closed-loop gains. **d**, Fluorescence transients during trial with low gain of 0.6 (Fly 15 from Fig. 1j). **e**, Comparison of PVA estimate versus accumulated rotation of visual cue and walking rotation on ball (trial in **d**). Walking rotation exceeds visual cue angular rotation in this low-gain trial. **f**, Similar to **d**, but with high closed-loop gain of 1.3 (Fly 3 from Fig. 1j). **g**, Similar to **e**, but with high gain (trial in **f**). **h**, Effective gain between walking rotation and PVA estimate for different closed-loop gains ($r = 0.69$, $p_r = 0$, Fit: slope $= 0.85 \pm 0.07$, $p_{slope} = 0$, $n = 172$, $R^2 = 0.48$, see Methods). **i**, Effective gain between visual cue rotation and PVA estimate for different closed-loop gains ($r = 0$, $p_r = 15.1 \times 10^{-3}$, Fit: slope $= -0.17 \pm 0.07$, $p_{slope} = 0.02$, $n = 172$, $R^2 = 0.03$, see Methods).

when the fly remained standing in a visual environment (Extended Data Fig. 10g–r), extending beyond durations that elicit adaptation in early visual circuits[29]. Thus, even in the absence self-motion cues, EBw.s population activity maintains a stable representation of the fly's orientation in its environment with or without visual landmarks.

Discussion

The ability of animals to combine continuous path integration with potentially intermittent landmark-based orienting enables navigation in a wide diversity of environmental conditions[1,6]. We studied the activity dynamics of a complete population of identified CX neurons in tethered walking flies and found that this network uses information from both landmark-based and angular path integration systems to create a compass-like representation of the animal's orientation in the environment.

Previous studies have described static visual maps in the CX[12,13,20]. Such maps may allow navigating insects to maintain a sun-compass-based heading direction[12,13,27,30]. Here we found that EBw.s neurons track the fly's orientation relative to visual landmarks in a variety of different visual environments (Figs 1 and 2), suggesting that the CX dynamically adapts to estimate the fly's orientation within its visual surroundings (Extended Data Figs 2d, 3d and 4g). Subsets of ring neurons are likely to bring information about spatially localized visual features[20] to specific rings of the ellipsoid body[31]. It is not yet clear how this information is converted into an abstract and flexible representation of the animal's orientation relative to landmarks[32], but EBw.s responses in a symmetric environment with two indistinguishable cues (Extended Data Figs 4 and 5) hint at an underlying winner-take-all process for landmark selection[33]. Combining landmark orientation with information about the animal's movement effectively creates an internal reference frame for the animal in its surroundings. Many of the proposed functions of the CX in directed locomotion[11,15], visual place learning[10], and action-selection[34], may rely on this internal reference. Although the EBw.s population tracks the fly's rotational movements in darkness, we do not yet know where and how translational motion, an important component of a complete navigational system, is incorporated. Additionally, although the calcium sensor we chose for our imaging experiments has the temporal resolution necessary to capture EBw.s representations of the fly's angular rotation (see Methods), it lacks the precision necessary for us to determine whether EBw.s activity represents the fly's predicted future orientation or its estimate of current orientation.

Our observation that EBw.s activity was maintained in the absence of self-motion suggests that internal dynamics play a significant role in shaping neural activity in the fly brain, much as they do in the brains of larger animals. Persistent activity in the CX can maintain compass information when the fly is standing in darkness for 30 s—two orders of magnitude longer than might be explained by calcium sensor decay kinetics[18]. Persistent activity has been shown to support maintenance of eye position in the goldfish[35] and has been proposed to underlie working memory in mammals[36]. In the CX, this activity may allow the fly to retain a short-term orientation memory even when landmarks are temporarily out of sight[8]. Consistent with this notion, the EBw.s activity bump largely remained tethered to the position of one landmark even in the presence of another identical landmark in front of the fly (Extended Data Fig. 4i). The bump also did not always shift instantaneously following an abrupt displacement of visual landmarks, as if temporarily retaining the original orientation reference before locking on to its new position (Extended Data Fig. 6b).

Several models have been proposed to explain how visual landmark and self-motion cues are integrated at the level of head direction cell activity in mammals[37]. Most rely on circuits organized as ring attractors: neurons are schematized as being arranged in a circle based on their preferred directions[38], with connection strengths that depend on their angular separation[37]. With initial sensory input and an appropriate balance of recurrent excitation and inhibition, such a circuit can generate and sustain a localized activity bump. The bump's position on the circle corresponds to the animal's heading which is then updated by directional drive from self-motion signalling neurons. Direct experimental evidence in support of these models has been difficult to obtain in mammals owing to the distributed nature of the underlying circuits. Although the functional connectivity between EBw.s neurons is

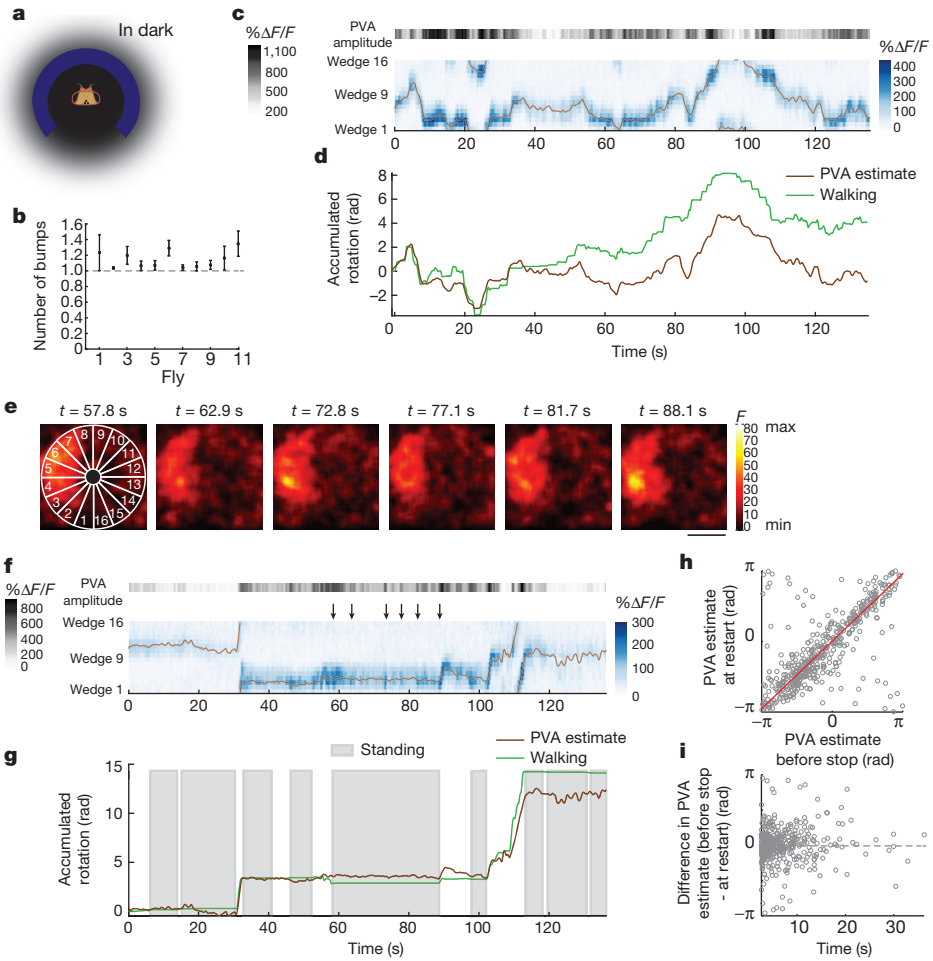


Figure 4 | Path integration, drift and persistence in EBw.s activity in total darkness. **a**, Experiments with flies walking in total darkness. **b**, Mean and s.d. of number of bumps across trials for each of 11 flies. **c**, Fluorescence transients during trial in darkness (Fly 9 in **b**). **d**, Accumulated ball rotation plotted against accumulated PVA estimate of fly's rotation. **e**, Sample frames from time series showing that EBw.s activity is maintained in absence of both visual cues and rotation (Fly 3 in **b**). Scale bar, 20 μm. **f**, Fluorescence transients during trial in **e**. **g**, Representation of fly's angular orientation is maintained in the absence of rotation and resumes from previous wedge after long delay (grey rectangles indicate epochs of fly standing). **h**, Comparison of PVA estimate of orientation before stop and at restart for different standing bouts across $n = 11$ flies ($r = 0.7$, $p_r = 0$, Fit: slope $= 0.96 \pm 0.17$, $p_{slope} = 0$, $n = 499$, $R^2 = 0.879$). **i**, Durations of standing bouts in **h** ($t_{mean} = 6.7 \pm 5.1$ s, $\Delta PVA_{mean} = 0.017 \pm 0.76$ rad).

not yet known, we have observed several of the expected features of ring attractor models[37,39,40] in the dynamics of this population of CX neurons: organization of activity into a localized bump, movement of the bump to neighbouring wedges based on self-motion, drift in bump location in darkness, persistent activity, and both abrupt jumps and gradual transitions of the activity bump when triggered by strong visual input. Cell-intrinsic mechanisms could also underlie some of these features, including, for example, persistent activity[35,41,42]. The genetic tools available in *Drosophila* to target and manipulate the activity of identified cell types should allow different models for visually guided orientation and angular path integration to be discriminated at the level of synaptic, cellular and network mechanism.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 December 2014; accepted 9 April 2015.

1. Collett, T. S. & Graham, P. Animal navigation: Path integration, visual landmarks and cognitive maps. *Curr. Biol.* **14,** R475–R477 (2004).
2. Mittelstaedt, M. L. & Mittelstaedt, H. Homing by path integration in a mammal. *Naturwissenschaften* **67,** 566–567 (1980).
3. Taube, J. S., Muller, R. U. & Ranck, J. B. Head-direction cells recorded from the postsubiculum in freely moving rats. 1. Description and quantitative analysis. *J. Neurosci.* **10,** 420–435 (1990).
4. Taube, J. S. The head direction signal: origins and sensory-motor integration. *Annu. Rev. Neurosci.* **30,** 181–207 (2007).
5. Huston, S. J. & Jayaraman, V. Studying sensorimotor integration in insects. *Curr. Opin. Neurobiol.* **21,** 527–534 (2011).
6. Wehner, R. Desert ant navigation: how miniature brains solve complex tasks. *J. Comp. Physiol. A* **189,** 579–588 (2003).
7. Collett, T. S. & Collett, M. Path integration in insects. *Curr. Opin. Neurobiol.* **10,** 757–762 (2000).
8. Neuser, K., Triphan, T., Mronz, M., Poeck, B. & Strauss, R. Analysis of a spatial orientation memory in *Drosophila*. *Nature* **453,** 1244–1247 (2008).
9. Liu, G. *et al.* Distinct memory traces for two visual features in the *Drosophila* brain. *Nature* **439,** 551–556 (2006).
10. Ofstad, T. A., Zuker, C. S. & Reiser, M. B. Visual place learning in *Drosophila melanogaster*. *Nature* **474,** 204–207 (2011).
11. Strauss, R. The central complex and the genetic dissection of locomotor behaviour. *Curr. Opin. Neurobiol.* **12,** 633–638 (2002).
12. Heinze, S. & Homberg, U. Maplike representation of celestial E-vector orientations in the brain of an insect. *Science* **315,** 995–997 (2007).
13. Heinze, S. & Reppert, S. M. Sun compass integration of skylight cues in migratory monarch butterflies. *Neuron* **69,** 345–358 (2011).
14. Pfeiffer, K. & Homberg, U. Organization and functional roles of the central complex in the insect brain. *Annu. Rev. Entomol.* **59,** 165–184 (2014).
15. Guo, P. & Ritzmann, R. E. Neural activity in the central complex of the cockroach brain is linked to turning behaviors. *J. Exp. Biol.* **216,** 992–1002 (2013).
16. Kathman, N. D., Kesavan, M. & Ritzmann, R. E. Encoding wide-field motion and direction in the central complex of the cockroach *Blaberus discoidalis*. *J. Exp. Biol.* **217,** 4079–4090 (2014).
17. Dombeck, D. A. & Reiser, M. B. Real neuroscience in virtual worlds. *Curr. Opin. Neurobiol.* **22,** 3–10 (2012).
18. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
19. Seelig, J. D. *et al.* Two-photon calcium imaging from head-fixed *Drosophila* during optomotor walking behavior. *Nature Methods* **7,** 535–540 (2010).
20. Seelig, J. D. & Jayaraman, V. Feature detection and orientation tuning in the *Drosophila* central complex. *Nature* **503,** 262–266 (2013).
21. Bausenwein, B., Muller, N. R. & Heisenberg, M. Behavior-dependent activity labeling in the central complex of *Drosophila* during controlled visual stimulation. *J. Comp. Neurol.* **340,** 255–268 (1994).
22. Strauss, R. & Pichler, J. Persistence of orientation toward a temporarily invisible landmark in *Drosophila melanogaster*. *J. Comp. Physiol. A* **182,** 411–423 (1998).
23. Jenett, A. *et al.* A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Rep.* **2,** 991–1001 (2012).
24. Wolff, T., Iyer, N. A. & Rubin, G. M. Neuroarchitecture and neuroanatomy of the *Drosophila* central complex: a GAL4-based dissection of protocerebral bridge neurons and circuits. *J. Comp. Neurol.* **523,** 997–1037 (2015).
25. Hanesch, U., Fischbach, K. F. & Heisenberg, M. Neuronal architecture of the central complex in *Drosophila melanogaster*. *Cell Tissue Res.* **257,** 343–366 (1989).

26. Lin, C. Y. *et al.* A comprehensive wiring diagram of the protocerebral bridge for visual information processing in the *Drosophila* brain. *Cell Rep.* **3,** 1739–1753 (2013).
27. Heinze, S., Gotthardt, S. & Homberg, U. Transformation of polarized light information in the central complex of the locust. *J. Neurosci.* **29,** 11783–11793 (2009).
28. Mizumori, S. J. Y. & Williams, J. D. Directionally selective mnemonic properties of neurons in the lateral dorsal nucleus of the thalamus of rats. *J. Neurosci.* **13,** 4015–4028 (1993).
29. Laughlin, S. B. The role of sensory adaptation in the retina. *J. Exp. Biol.* **146,** 39–62 (1989).
30. Bockhorst, T. & Homberg, U. Amplitude and dynamics of polarization-plane signaling in the central complex of the locust brain. *J. Neurophysiol.* http://dx.doi.org/10.1152/jn.00742.2014 (2015).
31. Young, J. M. & Armstrong, J. D. Structure of the adult central complex in *Drosophila*: organization of distinct neuronal subsets. *J. Comp. Neurol.* **518,** 1500–1524 (2010).
32. Zeil, J. Visual homing: an insect perspective. *Curr. Opin. Neurobiol.* **22,** 285–293 (2012).
33. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4,** 219–227 (1985).
34. Strausfeld, N. J. & Hirth, F. Deep homology of arthropod central complex and vertebrate basal ganglia. *Science* **340,** 157–161 (2013).
35. Aksay, E. *et al.* Functional dissection of circuitry in a neural integrator. *Nature Neurosci.* **10,** 494–504 (2007).
36. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Neurocomputational models of working memory. *Nature Neurosci.* **3,** 1184–1191 (2000).
37. Knierim, J. J. & Zhang, K. C. Attractor dynamics of spatially correlated neural activity in the limbic system. *Annu. Rev. Neurosci.* **35,** 267–285 (2012).
38. Peyrache, A., Lacroix, M. M., Petersen, P. C. & Buzsaki, G. Internally organized mechanisms of the head direction sense. *Nature Neurosci.* **18,** 569–575 (2015).
39. Arena, P., Maceo, S., Patané, L. & Strauss, R. A spiking network for spatial memory formation: towards a fly-inspired ellipsoid body model. *Intl. Joint Conf. Neural Networks* http://dx.doi.org/10.1109/IJCNN.2013.6706882 (2013).
40. Haferlach, T., Wessnitzer, J., Mangan, M. & Webb, B. Evolving a neural model of insect path integration. *Adapt. Behav.* **15,** 273–287 (2007).
41. Yoshida, M. & Hasselmo, M. E. Persistent firing supported by an intrinsic cellular mechanism in a component of the head direction system. *J. Neurosci.* **29,** 4945–4952 (2009).
42. Major, G. & Tank, D. Persistent neural activity: prevalence and mechanisms. *Curr. Opin. Neurobiol.* **14,** 675–684 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Wolff and G. Rubin for sharing information about CX neuron morphology. We thank Janelia Fly Core, and in particular K. Hibbard and S. Coffman, for support, J. Liu for technical support, and V. Iyer for ScanImage support. We are grateful to A. Karpova, A. Leonardo, S. S. Kim, H. Haberkern, D. Turner-Evans, C. Dan, S. Wegener and R. Franconville for discussions and comments on the manuscript. We thank W. Denk, S. Druckmann, J. Dudman, A. Lee, K. Longden, M. Reiser, S. Romani, G. Rubin, Y. Sun, and T. Wolff for feedback on the manuscript. This work was supported by the Howard Hughes Medical Institute.

Author Contributions Both authors designed the study, performed data analysis and wrote the manuscript. J.D.S. carried out all experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.J. (vivek@janelia.hhmi.org).

METHODS

Fly stocks. All calcium imaging experiments were performed with 8–11 days old female *UAS-GCaMP6f;R60D05-GAL4* flies. Flies were randomly picked from their housing vials for all experiments.

Nomenclature. EBw.s neurons are referred to variously as eb-pb-vbo[25], EIP[26] and EBw.s-PBg.b-gall.b[24] neurons in the fly literature. They may be homologous to CL1a neurons in the locust[27] and butterfly[13].

Fly preparation for imaging during walking. The fly was anaesthetized on ice and transferred to a cold plate at 4 °C. The fly's proboscis was pressed onto its head and immobilized with wax. To maximize the fly's field of view we used a two-piece pyramidal stainless steel shim holder[19] similar to those previously used for tethered flying fly experiments[20,43]. The fly was glued to a pin and positioned in the holder using a micromanipulator and fixed in the holder with UV gel as previously described[19,20]. The fly body axis was angled at 31° ± 5° (measured for 5 flies) with respect to the tracking system to orient the EB optimally with respect to the microscope's focal plane. To stop brain movement due to the pulsation of muscle M16, we cut the muscle—or the nerves innervating the muscle—with dissection needles if necessary. The fly holder (including the micromanipulator) was then transferred to the two-photon microscope and secured using magnetic mounts. As previously described[19,20], the fly was positioned on an air-supported ball with a three-axis micromanipulator and the walking velocity of the fly was monitored using a camera system. For all experiments described in the main figures we used a 6-mm diameter, 40-mg ball[19]. For the experiments in the dark described in Extended Data Fig. 7h–j, we used a 10-mm diameter, 175-mg ball. All balls were made of polyurethane foam.

For experiments with visual cues, we removed parts of the antennae (funiculus and arista) to reduce the fly's tendency to touch the holder.

For experiments in which flies walked in the dark we additionally coated the eyes of a subset of flies with black paint (Premiere Acrylic Colours, Mars Black). For the 6-mm-ball experiments, Flies 4–11 had coated eyes, while Flies 2–13 had coated eyes for the 10-mm-ball experiments. The number of trials per fly was as follows (Fly (number of trials)). 6-mm ball: 1(15), 2(8), 3(8), 4(8), 5(11), 6(10), 7(8), 8(10), 9(17), 10(10), 11(11). 10-mm ball: 1(7), 2(3), 3(8), 4(12), 5(10), 6(8), 7(6), 8(5), 9(3), 10(14), 11(8), 12(6), 13(11). All trials across all conditions lasted 140 s.

Two-photon calcium imaging. Calcium imaging was performed using a custom-built two-photon microscope controlled with ScanImage 4.2[44]. We used an Olympus ×40 objective (LUMPlanFl/IR, NA 0.8) and a GaAsP photomultiplier tube (H7422PA-40, Hamamatsu). A Chameleon Ultra II laser (Coherent, Santa Clara, CA) tuned to 920 nm was used as the excitation source with the power adjusted to below 20 mW at the sample. We used the same saline as in previous studies[20] but adjusted the calcium concentration to 2.5 mM. Focal planes were selected to optimize coverage of the part of the EB innervated by EBw.s neurons. We imaged from 5-plane volumes at a rate of 8.507 Hz with an equal spacing of between 4 μm to 6 μm between individual focal planes.

The calcium signals we measured may reflect synaptic input, action potential output or some combination of both. We chose GCaMP6f[18] for our experiments because it offers the temporal resolution necessary to capture EBw.s representation of the fly's angular rotation. Based on *in vivo* measurements of responses to 20 Hz spiking at the *Drosophila* larval neuromuscular junction, GCaMP6f has a time-to-peak of ~141 ms (close to the 8.507 Hz frame rate of our imaging system) and a decay time of ~380 ms (ref. 18). Assuming that one complete rotation of the fly is represented by activity moving through the 16 wedges of the EB, each wedge represents 22.5° of rotation. The maximum average rotational velocity reached by a fly in our experiments was ~ 35° s^{-1} (Extended Data Fig. 1), which would result in a bump of activity moving across a wedge no faster than ~640 ms on average. Thus, possible lags in the calcium signals introduced by the rise and decay times of GCaMP6f would not compromise the detection of these activity changes. Although we do not know the actual change in electrical activity associated with the calcium transients we see, the kinetics of GCaMP6f provide a considerable margin of error.

Visual stimulation. *Visual arena.* Visual stimuli were presented on a cylindrical LED display[45] spanning 270° in azimuth and 120° in elevation, and tilted by 10° towards the fly. The display was covered with a colour filter and a diffuser as previously described[19,20].

Visual stimuli for closed-loop walking experiments. We used three different visual stimuli: condition 1, a bright vertical stripe spanning 120° in elevation and 15° in azimuth; condition 2, two bright stripes of the same dimensions separated by 135° (resulting in a pattern that was invariant to rotations by 135°); and condition 3, a pattern containing several vertical and horizontal stripes (Extended Data Fig. 1a–c). The number of trials (in brackets) for each fly for each of these conditions was: trial 1, 1(7), 2(6), 3(9), 4(11), 5(15), 6(13), 7(3), 8(7), 9(7), 10(6), 11(5), 12(9), 13(9), 14(10), 15(6); trial 2, 1(12), 2(8), 3(10), 4(3), 5(6), 6(8), 7(16); trial 3, 1(6) (same as fly 1 in trial 2), 2(2) (same as fly 2 in trial 2), 3(5) (same as fly 4 in trial 2), 4(3), 5(7) (same as fly 5 in trial 2), 6(11), 7(4), 8(11), 9(12).

In cue-shift experiments each trial consisted of two cue shifts by the same angular distance within each trial (either by 60° or 120°) after at least 50 s of closed-loop behaviour. The first cue shift was counter clockwise from the current position and the next, 50 s later, clockwise by the same angular amount from the current position. Cue shift experiments were performed with a subset of the flies in condition 1. A 60° cue shift was used for flies 7(2), 11(3), 13(3), 15(2) and 120°cue shift for flies 7(2), 8(3), 9(1), 11(3), 13(3), 15(2).

In experiments that tested the influence of prior exposure to visual stimuli in closed loop on the gain between walking rotation and PVA estimate in darkness (Extended Data Fig. 9), we exposed the flies to 65 s of closed-loop walking with either low gain (mean closed-loop gain = 0.47 ± 0.04, close to the fly's default gain on the ball without prior closed-loop walking experience) or higher gain (mean set gain = 0.9 ± 0.16) with a single stripe, after which the stripe disappeared and the trial continued for another 60 s in darkness. For these experiments, we only used flies that showed strong rotational movement with low drift—as assessed at the onset of the experiment—to increase the accuracy of the gain calculation. We only recorded a small number of trials per fly, because the fly usually rotated more at the onset of experiments and walked forward more towards the end, leading to increased drift. For a subset of flies, we also tested the intrinsic gain of the flies walking in darkness without prior exposure to the closed-loop stimulus. The number of trials for experiments in which we combined closed-loop walking and walking in darkness were (fly number (trials with disappearing stripe/trials in darkness before exposure to closed-loop condition)): 1(5/0), 2(4/1), 3(3/1), 4(5/0), 5(7/1), 6(3/2), 7(4/1), 8(6/2), 9(5/1), 10(4/1), 11(4/1), 12(3/2), 13(5/2), 14(5/1), 15(1/2), 16(5/2), 17(3/2), 18(6/2), 19(4/2), 20(4/2), 21(5/2), 22(4/2), 23(4/1), 24(2/2), 25(4/2), 26(3/2).

Closed-loop gains to convert rotation on the ball to displacement of the stripe around the arena were close to 1 ('normal gain'), but were varied from 0.4 to 1.6 in experiments to explore the effect of gain change on EBw.s representation. Actual values of the gain were verified by fitting changes in ball displacement to changes in pattern displacement on the arena. All patterns were displaced directly from one edge of the 270° arena to the other behind the fly rather than having them progress virtually through the 90° of visual field not represented by the arena. This was done to prevent abrupt changes in light intensity and to keep the number of features in the fly's visual field constant.

All experiments with visual stimulation were performed in closed loop[46]. The voltage position signal of the tracking system was read with a DAQ board and discretized in 20-ms intervals using custom LabVIEW software which was also used to update the position of the visual stimulus[45].

Data analysis. We used MATLAB (MathWorks, Inc., Natick, MA) and the Circular Statistics Toolbox[47] for data analysis. All errors and error bars shown are standard deviation (s.d.). No statistical methods were used to predetermine sample size.

Calculation of fluorescence changes. Each imaged volume (stack of five frames) was averaged for analysis—we refer to this average as a 'frame'. Each frame was spatially filtered with a 2-pixel-wide Gaussian filter after which background fluorescence was subtracted. Calcium transients recorded from behaving flies were smoothed with a 3rd order Savitzky–Golay filter over 7 frames (822 ms) for comparisons with behavioural data. The baseline for calculating $\Delta F/F$ was computed by averaging over the 10% of lowest-intensity frames in each trial. For display only, MIP fluorescence intensity images shown in Figs 1e, f, 2a, 4e, and Extended Data Fig. 5a were filtered with a 20-pixel-wide, 10-pixel-s.d. Gaussian filter (the size of each image is 216 pixels by 216 pixels).

ROI selection. ROIs corresponding to 16 wedges of the EB were selected manually in videos of $\Delta F/F$ by drawing an ellipse (with a central hole, as depicted in Fig. 1f, g) that surrounded the EB, and then equally subdividing the ellipse into 16 wedges each spanning 22.5°. The number of wedges was selected based on the well-characterized EB wedge and PB glomerular innervation patterns of EBw.s neurons labelled by the *R60D05-Gal4* line[24]. Some EBw.s neurons are known to arborize only in half- or demi-wedges[24]. Thus, our ROI selection and population analysis strategy may underestimate the actual resolution of the EBw.s system.

Population vector average (PVA). The PVA was computed as the weighted average of EB wedge angles ranging from 0 to 360°, with average $\Delta F/F$ values for each wedge used as a weight. This PVA estimate was smoothed with a box-car filter over 3 frames (352 ms). We used brewermap (S. Cobeldick, MathWorks file exchange) with colour schemes from http://colorbrewer2.org/ to generate colour maps for all PVA plots except for PVA amplitude, which we display in greyscale. For display of PVA estimates of orientation or walking rotation, raw PVA was offset by the median difference (circular distance) between PVA and either the visual cue position (for closed-loop trials in the arena) or the walking rotation signal (for trials in the dark). We computed the offset using epochs of walking in

the final 80% of a trial, a period during which PVA estimates were typically more stable. The offset adjustment was necessitated by the fact that there was no stereotyped relationship between cue positions and EBw.s signal across flies (Fig. 1m, Extended Data Fig. 2e, f, Fig. 2f, Extended Data Figs 3e, f, 4j, l, m). The magnitude of the offset in many animals (Fig. 1m, Extended Data Figs 2e, f, 3e, f, 4j, l) greatly exceeded the slight variance in the angle at which the tethered fly's head was fixed relative to the LED arena. The offset also occasionally changed between trials for the same fly. We did not monitor the fly's walking between trials, leaving open the possibility that these differences in offset arose purely from rotational movements of the fly (in the absence of closed-loop visual feedback) before initiation of the next closed-loop trial.

Analysis of number and width of activity bumps. All ROIs with calcium transients above a set threshold were included in an activity bump. Each contiguous set of ROIs above threshold defined an individual bump. We used two separate methods to set the threshold. In Method 1 (used for all the main figures), the threshold was defined as 1-s.d. above the mean of calcium transients across ROIs for each imaging frame. In Method 2 (see Extended Data Figs 2, 3, 4 and 7), the threshold was defined as the mean of calcium transients across ROIs over the entire trial. The width of a bump in each frame was, in all cases, defined as the full width at half maximum (with minimum in each frame subtracted). We used the Kolmogorov–Smirnov two-sample test for tests of the null hypothesis that bump widths for two different stimulus conditions are drawn from the same distribution. *P* values for this test are shown in the relevant figure legends.

Offset between pattern position and PVA estimate. The offset between the pattern position and the PVA estimate was calculated as the circular distance between the PVA and the leftmost pixel of the pattern in Extended Data Fig. 1a–c across the entire trial for Extended Data Figs 2e, 3e and 4l, and averaged over all trials for each fly in Figs 1m and 2f and Extended Data Fig. 4j. The s.d. of the offset was calculated as the circular s.d.[47] of the offset signal in each trial, and averaged across all trials. The pattern position from 0° to 270° was mapped to 0° to 360°, as explained below.

Mapping of 270° visual arena onto 360° EB. To compute the mapping of the visual pattern onto the EB we calculated the gain—slope of a linear fit—between the unwrapped (see below) pattern position and the unwrapped PVA estimate for those trials in which the pattern moved over at least half the display. Since EBw.s activity in response to cues on the 270° LED arena was uniformly mapped to 360° of the EB (Extended Data Figs 2d, 3d, 4g), visual cue positions on the 270° arena were mapped linearly to an arena spanning 360° in all plots and analyses to facilitate comparisons of cue position with PVA estimates and walking rotation.

Walking behaviour analysis. Ball movement was recorded at a sampling rate of 4 kHz[19]. Ball displacement and stimulus position were down-sampled to match the corresponding two-photon scan rate (8.507 Hz). Velocities in Extended Data Fig. 1 were calculated over 20 frames (2.35 s) and averaged over the entire trial. Walking traces were subdivided into walking and standing epochs—only epochs that lasted at least 20 frames were considered for such classification. We only labelled epochs as 'standing' if the fly was standing for at least 20 successive imaging frames (2.35 s).

Correlation analysis. Pearson's correlation coefficients were computed between two entire 'unwrapped' time series, which is the cumulative sum of all angular displacements. For display only, we 'wrapped' the angular data into the $-\pi$ to π range. Only trials in which the fly walked for more than 30% of the time were used in summary plots of correlations and gains. For Fig. 1, we can reject the null hypothesis that true single-trial correlations are 0 with $P < 0.05$ for all trials except for 1/12 trials of Fly 3, 1/15 trials of Fly 5 and 1/9 trials of Fly 13. For these trials, the correlations are 0.028, 0.053 and 0.024 respectively. For Fig. 2g (multiple features) and Extended Data Fig. 4k (two stripes), *P* values for correlations for all trials of all flies are <0.05. For Extended Data Fig. 7b (walking in darkness), $P > 0.05$ for the correlations for only the following trials: 1/8 trials of Fly 3, 1/8 trials of Fly 4, 1/9 trials of Fly 6, 1/10 for Fly 8, 1/7 for Fly 9, 1/10 for Fly 10. For these trials, correlations are 0.056, 0.037, −0.023, −0.054, 0.039, and −0.026 respectively.

Computation of gains. Closed-loop gains that translated ball rotation into movements of visual patterns on the LED arena were set to fixed values for each trial. However, differences in infrared lighting conditions affected the optical mouse sensor chip system's computation of ball rotation slightly[19], resulting in small variations in effective closed-loop gains. To compute true gain values, the ball's rotation about the vertical axis was linearly fit against the pattern position. The slope of this line was considered the actual gain for the trial.

The relationship between rotation of activity around the EB and either walking rotation or visual cue rotation on the arena was captured by a linear fit. The gain between EBw.s activity rotation and either behaviour or stimulus was computed as the slope of this line.

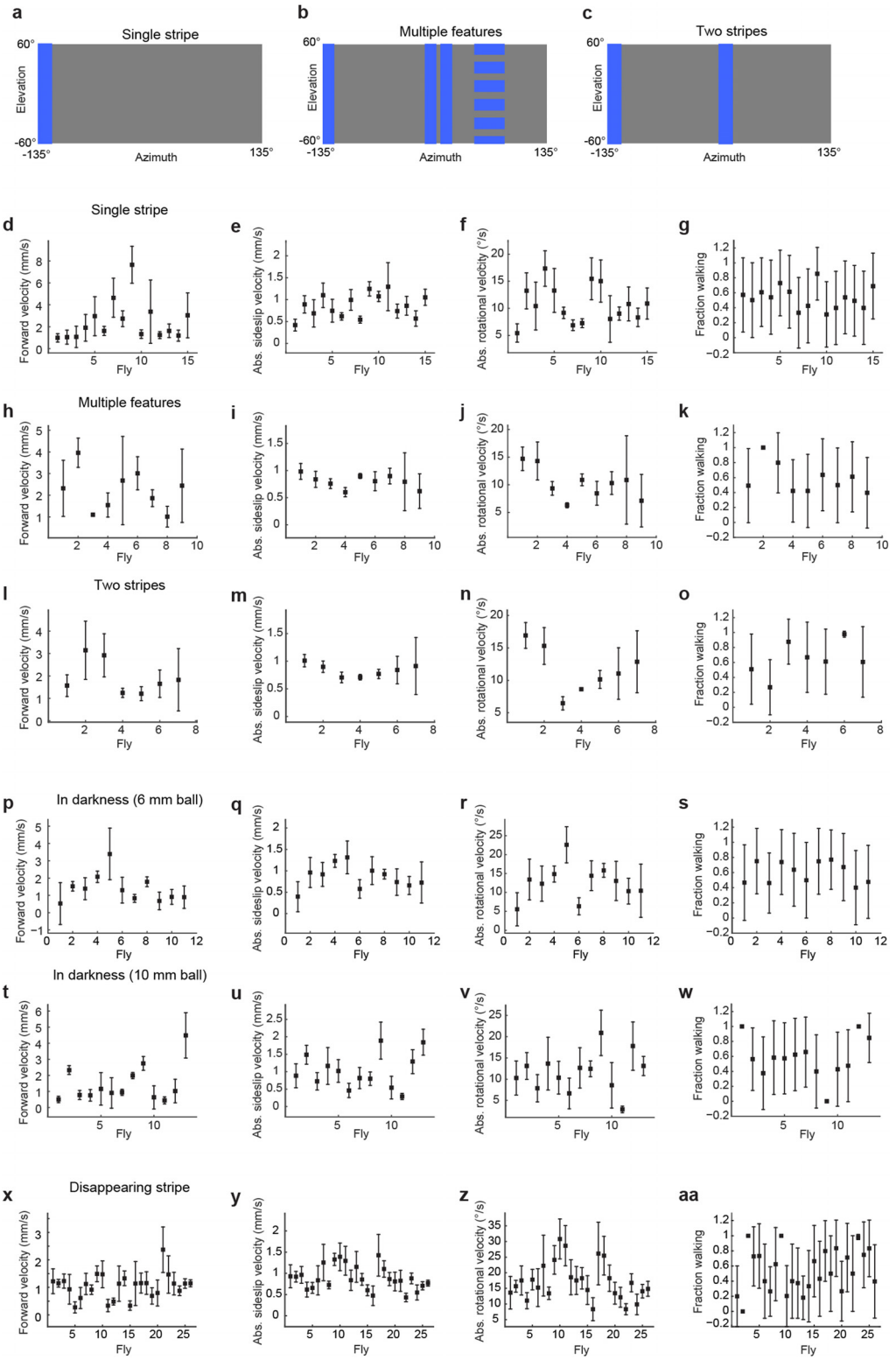
In all cases above, gains were computed across 200 frames (23.51 s) or over the entire walking epoch for data in Fig. 3h, i. Two-dimensional distributions of correlation values for flies walking in darkness were computed using a window of 200 frames sliding along the time series in steps of 25 frames (2.94 s). For Fig. 3h, i we only included walking epochs for which the visual cue moved over at least half of the display, and calculated gains over the entire walking bout.

Comparison of angular velocity with PVA-estimated velocity. For Extended Data Fig. 8, we computed angular velocity and PVA-estimated angular velocity using a 20-frame window (2.35 s) and plotted the values against each other for all trials in darkness for each fly. Points were then coloured based on the mean PVA amplitude during the 20-frame epoch.

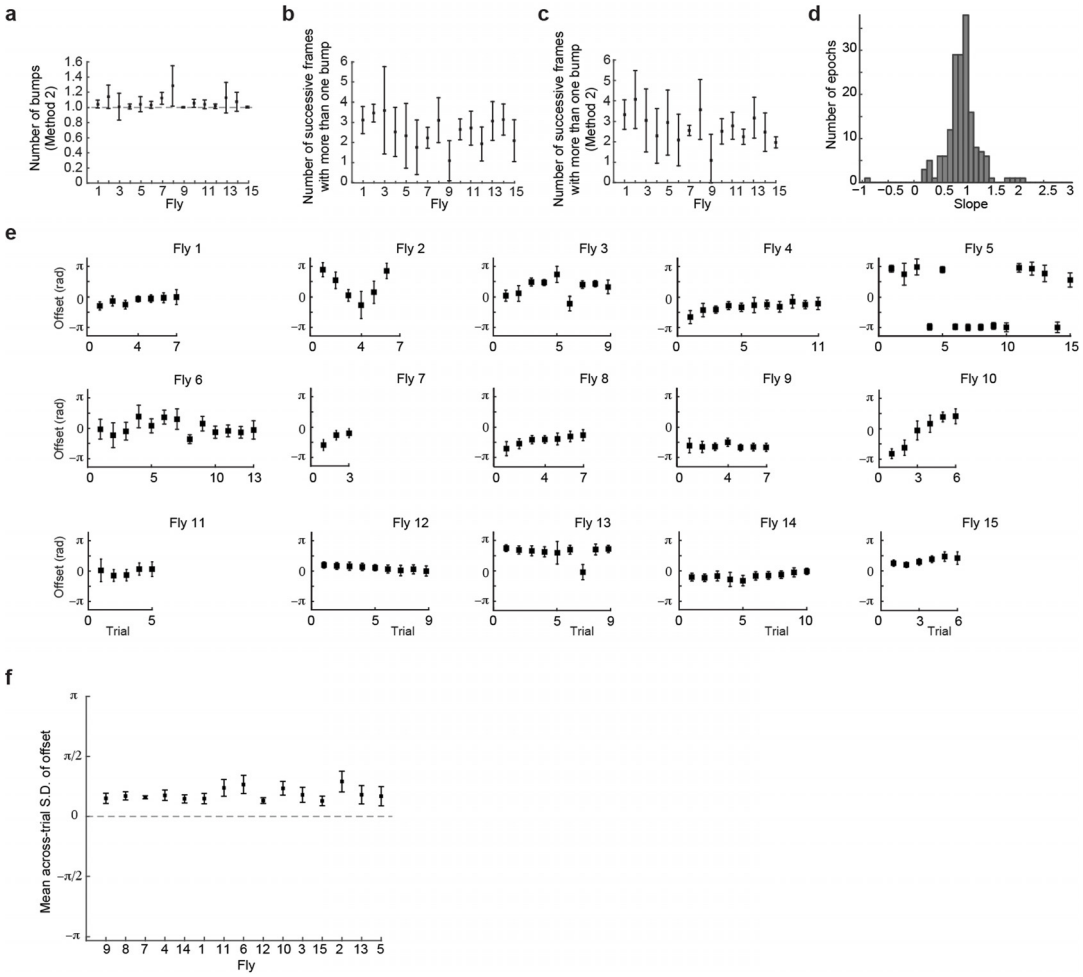
Analysis of persistent activity. To compare changes in PVA estimates during periods when the fly was not walking, we selected epochs of alternating walking-standing-walking bouts, with walking bouts each lasting at least 5 frames (588 ms) and non-walking bouts lasting at least 20 frames (2.35 s), well beyond the persistence of calcium signals attributable to the decay kinetics of the indicator. All values were averaged over 5 frames (588 ms) before or after the stop and restart of walking, respectively.

Analysis of responses to cue shifts. Changes in the offset between the visual cue position (the leftmost pixels of the cue seen from the fly's perspective) and the PVA estimate were computed as the mean circular distance over 100 frames (11.76 s). We compared the 100-frame mean offset before the first cue jump to the offset before the second cue jump, and the offset before the second cue jump to the offset at the end of the trial. For comparison, we also computed the expected change in PVA-cue offset if the PVA were not to follow visual cue position, in which case the PVA-cue offset would change by the magnitude of the cue jump.

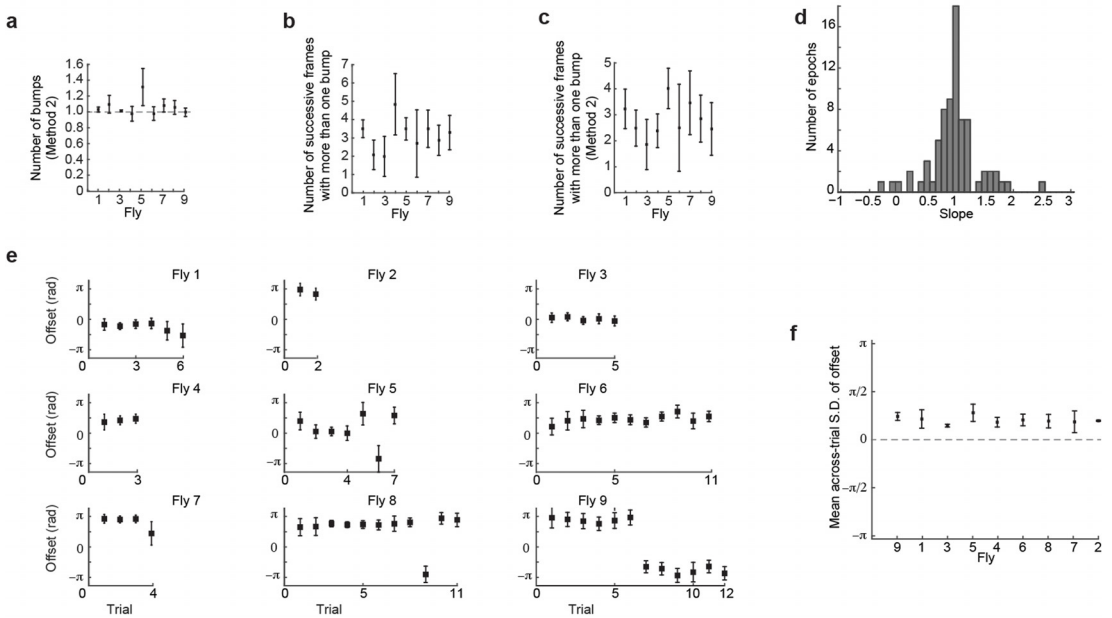
43. Maimon, G., Straw, A. D. & Dickinson, M. H. Active flight increases the gain of visual motion processing in *Drosophila*. *Nature Neurosci.* **13,** 393–399 (2010).
44. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2,** 13 (2003).
45. Reiser, M. B. & Dickinson, M. H. A modular display system for insect behavioral neuroscience. *J. Neurosci. Methods* **167,** 127–139 (2008).
46. Bahl, A., Ammer, G., Schilling, T. & Borst, A. Object tracking in motion-blind flies. *Nature Neurosci.* **16,** 730–738 (2013).
47. Berens, P. CircStat: A MATLAB toolbox for circular statistics. *J. Stat. Softw.* **31,** 1–21 (2009).



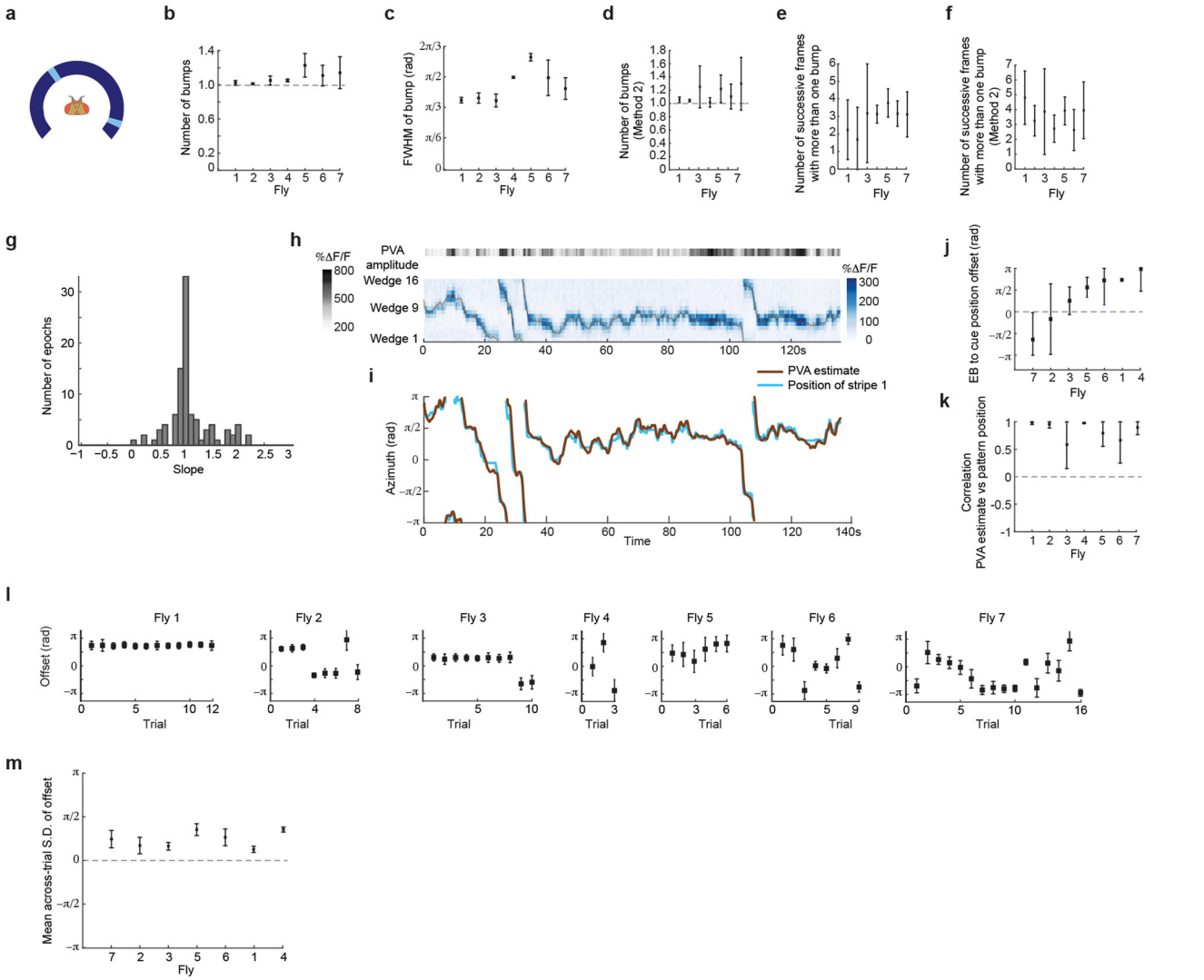
Extended Data Figure 1 | Visual stimuli, walking velocities and fraction of time walking across flies and conditions. **a**, Single-stripe pattern. **b**, Pattern with multiple features. **c**, Pattern with two identical stripes positioned symmetrically on the 270° visual display. In all closed-loop experiments, visual stimuli wrapped around the 270° arena, going directly from 0° to 270° and vice versa. **d**–**g**, Walking performance during closed-loop walking with a single stripe: **d**, forward velocity; **e**, magnitude of sideslip velocity; **f**, magnitude of rotational velocity; **g**, fraction of time walking across all trials. **h**–**k**, Same as **d**–**g** for the pattern with multiple features. **l**–**o**, Same as **d**–**g** for pattern with two stripes. **p**–**s**, Same as **d**–**g** for walking in the dark on a 6-mm diameter ball. **t**–**w**, Same as **d**–**g** for walking in the dark on a 10-mm diameter ball. **x**–**aa**, same as **d**–**g** for experiments with trials that combined epochs of closed-loop walking with epochs of walking in darkness (Extended Data Fig. 9).



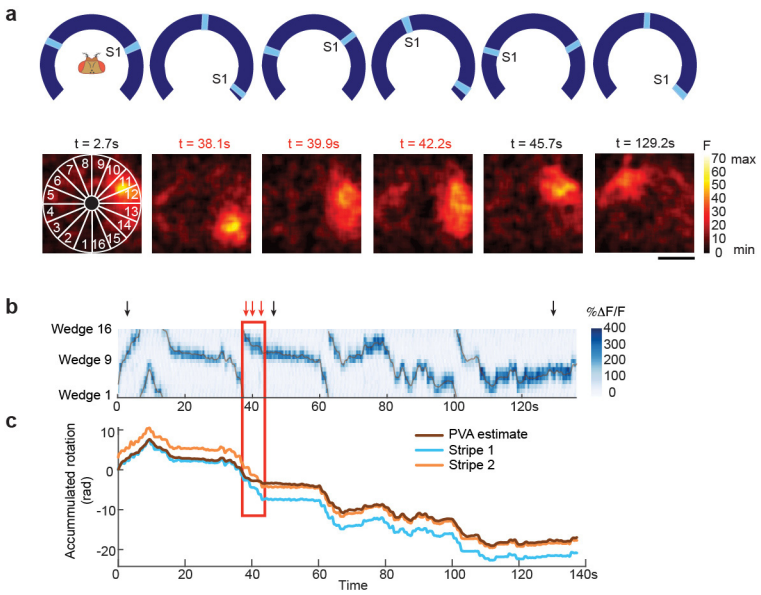
Extended Data Figure 2 | Closed-loop walking in visual environment with single stripe pattern. **a**, Mean and s.d. of the number of activity bumps as measured by Method 2 (see Methods) during all trials of all flies shown in Fig. 1. **b**, Mean and s.d. of the number of successive calcium imaging frames (recorded at 8.507 Hz) with more than one bump, measured using Method 1 (see Methods), for all flies shown in Fig. 1. **c**, Same as **b**, but computed using Method 2. **d**, Histogram of slopes of the linear fit between PVA estimate and pattern position during walking epochs, that is, the gain between unwrapped PVA estimate and unwrapped pattern position. The pattern was mapped from 0°-to-270° to 0°-to-360° for PVA calculations (see Methods). Thus, a slope of 1 corresponds to a visual pattern on the 270° arena that maps to the entire ring of the ellipsoid body. Only those walking epochs during which the pattern moved over at least half of the visual display were included so as to obtain an accurate estimate of the slope (mean slope = 0.92 ± 0.32, $n = 172$ walking epochs, see Methods). **e**, Mean and s.d. of angular offsets between PVA position and pattern position for each trial (140 s, see Methods) for all flies. **f**, Mean and s.d. of s.d. of angular offset between PVA position and pattern position.



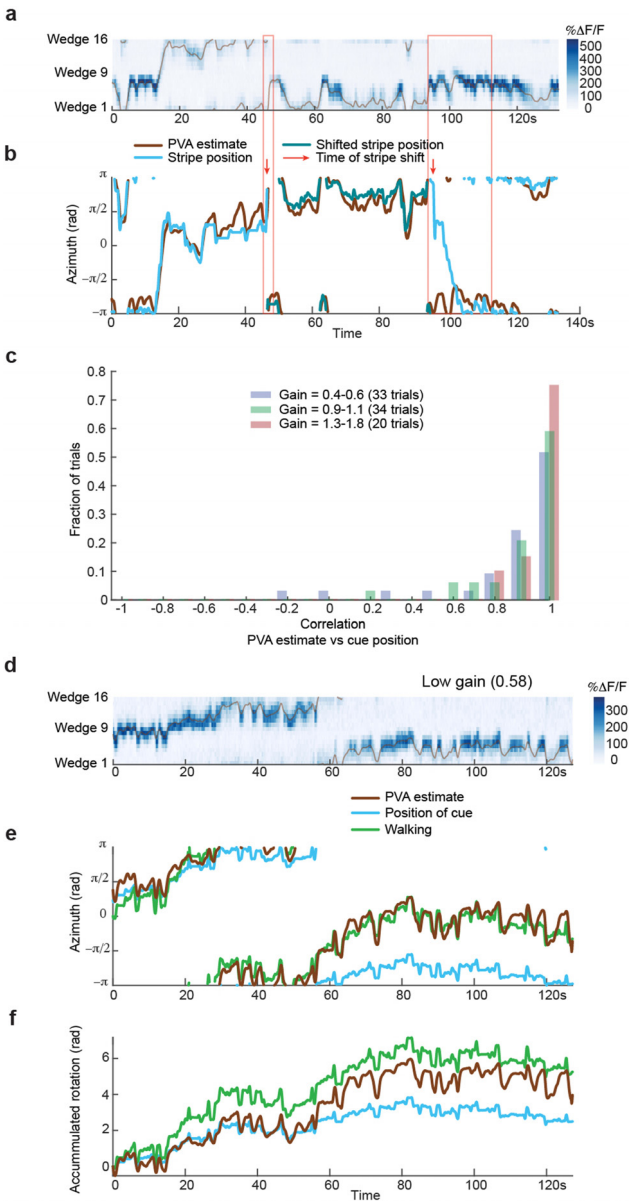
Extended Data Figure 3 | Closed-loop walking in visual environment with multiple features. **a**, Mean and s.d. of the number of activity bumps as measured by Method 2 (see Methods) during all trials of all flies shown in Fig. 2. **b**, Mean and s.d. of the number of successive calcium imaging frames with more than one bump, measured using Method 1 (see Methods), for all flies shown in Fig. 2. **c**, Same as **b**, but computed using Method 2. **d**, Same as Extended Data Fig. 2d for the pattern with multiple features (mean slope = 0.97 ± 0.43, $n = 74$ walking epochs). **e**, Mean and s.d. of angular offsets between PVA position and pattern position for each trial (140 s) for all flies. **f**, Mean and s.d. of s.d. of angular offset between PVA position and pattern position.



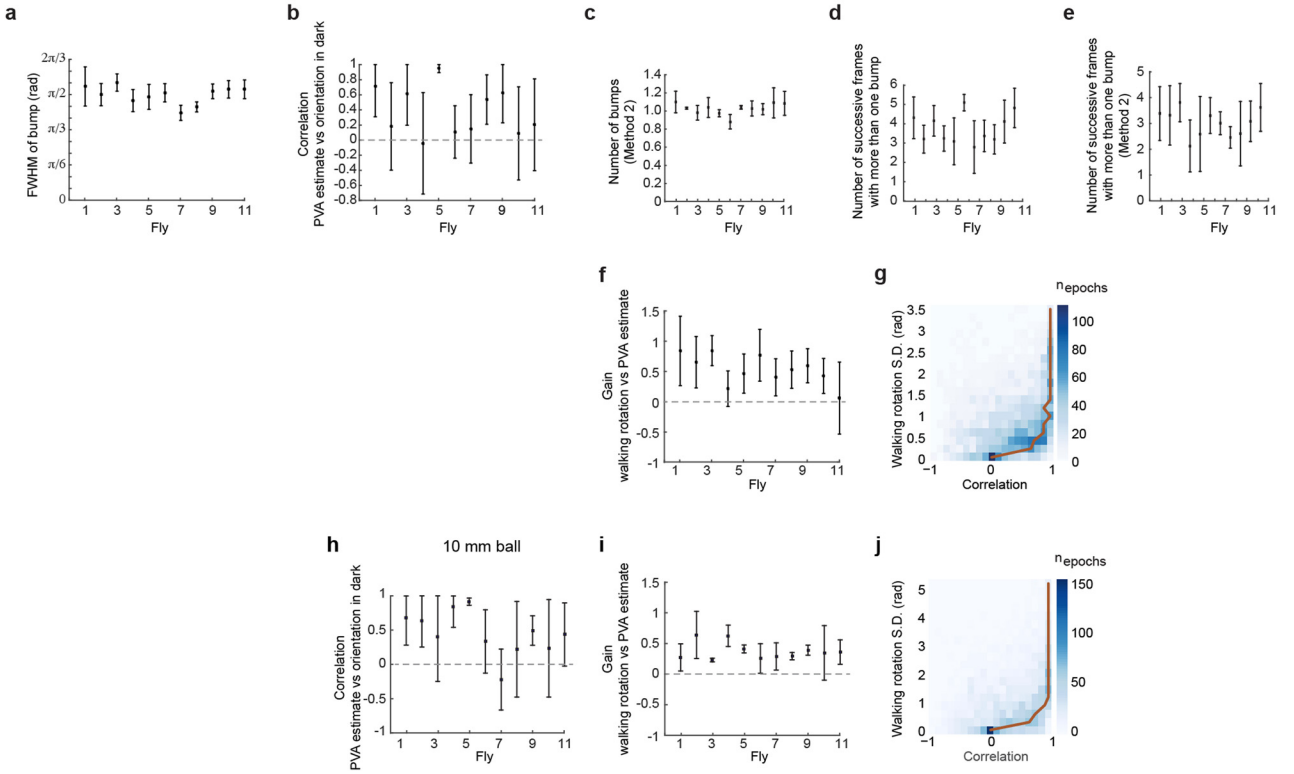
Extended Data Figure 4 | Single activity bump during closed-loop walking in visual environment with two stripes. **a**, Closed-loop experiment in visual environment with two identical and symmetrically placed stripes. **b**, Mean and s.d. of number of bumps in EBw.s population activity across trials for each of 7 flies. **c**, Mean and s.d. of FWHM of bump. Distribution of bump widths is significantly different from that for single-stripe stimulus (Fig. 1k); $P = 4.5 \times 10^{-6}$ (see Methods), mean width $= 78.7^{\circ} \pm 15.6^{\circ}$ for two-stripe trials versus $82.3^{\circ} \pm 11.5^{\circ}$ for single-stripe trials. **d**, Mean and s.d. of the number of activity bumps as measured by Method 2 (see Methods) during all trials for all flies. **e**, Mean and s.d. of the number of successive calcium imaging frames with more than one bump, measured using Method 1 (see Methods). **f**, Same as **e**, but computed using Method 2. **g**, Same as Extended Data Fig. 2d for the pattern with two stripes (mean slope $= 1.08 \pm 0.41$, $n = 96$ walking epochs). **h**, EBw.s fluorescence transients during trial with two-stripe pattern (Fly 2 in **b**). **i**, PVA estimate of the fly's angular orientation compared to actual orientation. **j**, Mean and s.d. of angular offsets between PVA position and pattern position in all flies. **k**, Correlation between PVA estimate and actual orientation of original left stripe for all flies. **l**, Mean and s.d. of angular offsets between PVA position and pattern position for each trial for all flies. **m**, Mean and s.d. of s.d. of angular offset between PVA position and pattern position.



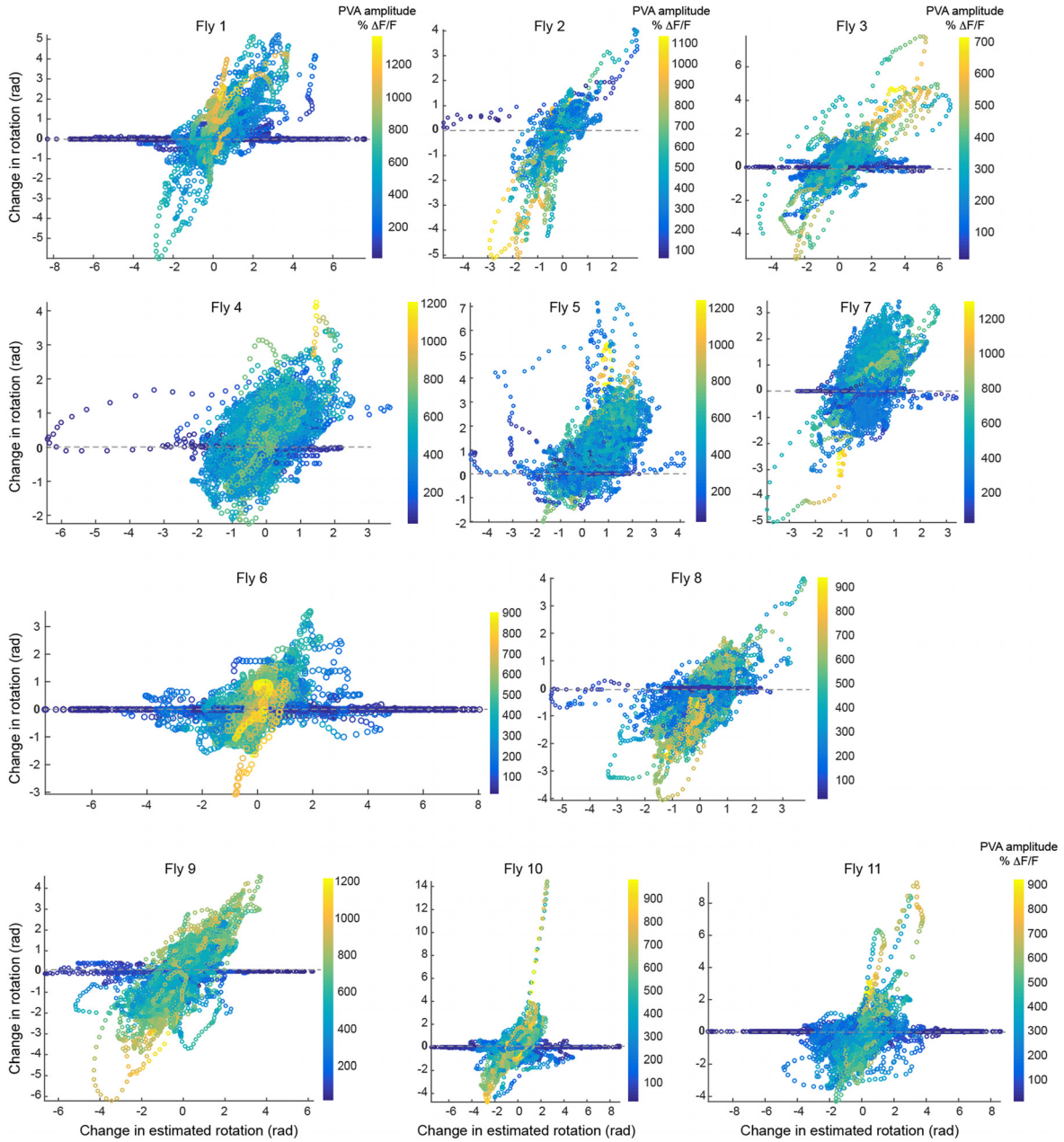
Extended Data Figure 5 | Example of EBw.s activity bump transitioning between locking to one of two identical visual cues placed symmetrically on LED arena. **a**, Sample frames from a calcium imaging time series showing single bump of EBw.s activity as the two-stripe pattern moved around the arena in a trial in which correlation between EBw.s activity and PVA estimate changes over a 4-s period (Fly 6 in Extended Data Fig. 4b). Frames during jump indicated by red time stamps. Scale bar, 20 μm. **b**, EBw.s fluorescence transients during trial displayed in **a**. **c**, Decoding of fly's angular orientation using unwrapped PVA of EBw.s activity plotted against the fly's unwrapped orientation with respect to stripe 1 and stripe 2 in the visual scene with two stripes. Red box corresponds to period when the EB activity bump switches from locking to one stripe to locking to the other (identical) stripe.



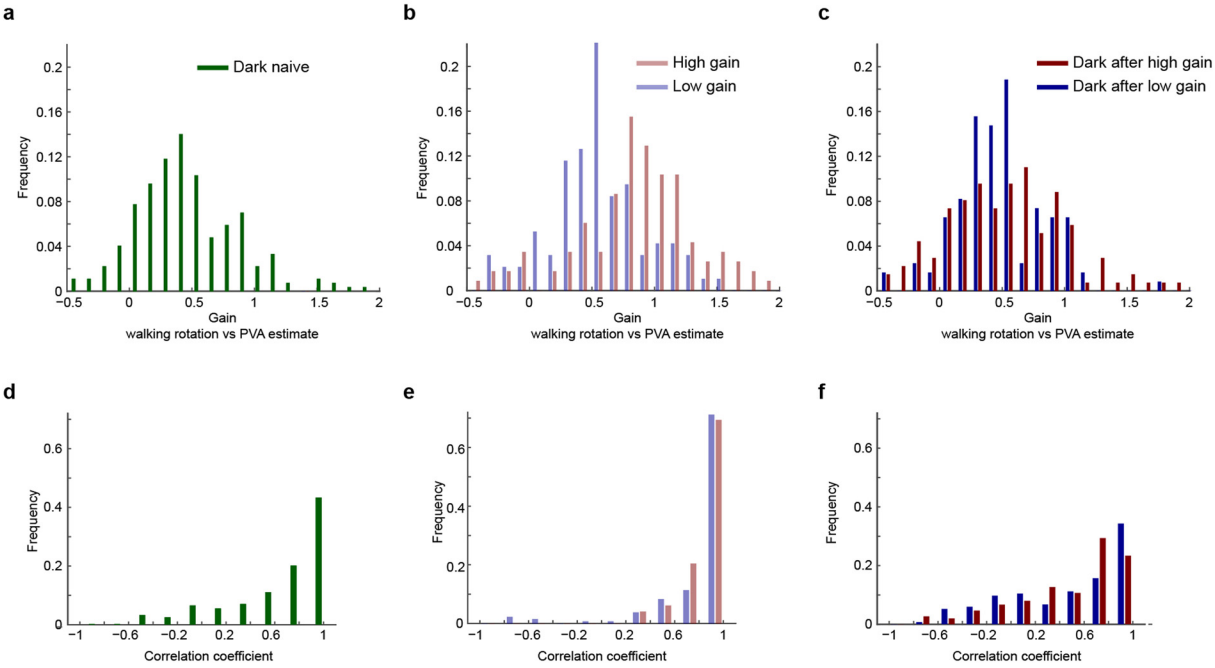
Extended Data Figure 6 | Competing influences of visual cue and self-motion on EBw.s activity. **a**, Fluorescence transients during cue shift trial (Fly 9 from Fig. 1j). Red box highlights epochs during which cue abruptly shifted to new position. **b**, Comparison of PVA estimate versus actual orientation. **c**, Correlations between PVA estimates and actual orientation relative to visual cue across trials and flies for different closed-loop gain values. **d**, Fluorescence transients in the EB during closed-loop trial with a low gain of 0.58 (Fly 6 in Fig. 1j-m). Superimposed brown line indicates PVA estimate of orientation. **e**, Decoding of fly's angular orientation using PVA of EBw.s activity plotted along with the pattern position and the fly's walking rotation. PVA closely matches walking rotation rather than visual cue rotation. Note that walking rotation exceeds visual cue angular rotation in this low gain trial. **f**, Comparison of PVA estimate versus accumulated rotation of visual cue and accumulated walking rotation on the ball shows PVA estimate more closely matches walking rotation than visual cue rotation.



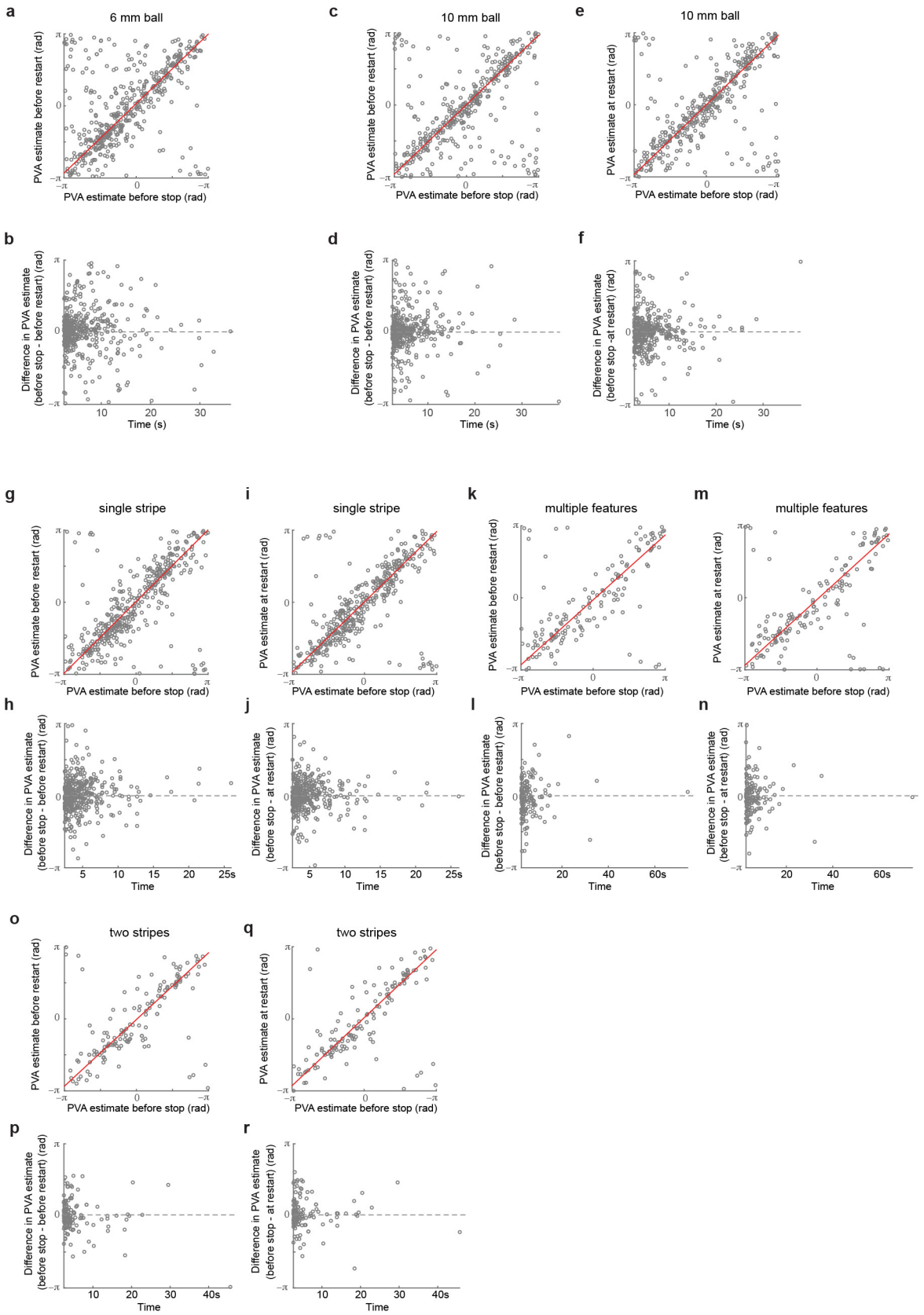
Extended Data Figure 7 | EBw.s activity when flies walk in darkness on balls of two different diameters. **a**, Mean and s.d. of FWHM of bump when walking in darkness on 6-mm ball. Distribution of bump widths is significantly different from that for single-stripe stimulus (Fig. 1k); $P = 8 \times 10^{-9}$ (see Methods), mean width $= 90.9^{\circ} \pm 11.2^{\circ}$ for walking in darkness versus $82.3^{\circ} \pm 11.5^{\circ}$ for single stripe. **b**, Correlations between accumulated PVA and walking rotation in the dark across flies for walking on 6-mm diameter ball. **c**, Mean and s.d. of the number of activity bumps as measured by Method 2 (see Methods) during all trials (6-mm ball). **d**, Mean and s.d. of the number of successive calcium imaging frames with more than one bump, measured using Method 1 (see Methods, 6-mm ball). **e**, Same as **d**, but computed using Method 2 (6-mm ball). **f**, Gain between accumulated PVA estimates of orientation and accumulated walking rotation across flies for 6-mm ball. **g**, Sliding window correlations (200 frames with a step size of 25 frames) between accumulated PVA estimate and accumulated walking rotation for different levels of s.d. of walking rotation for 6-mm ball (s.d. cutoff shown included 97% of epochs). Brown line connects highest-frequency bins. **h**, Correlations between accumulated PVA and walking rotation across flies when walking in the dark on 10-mm diameter ball. **i**, Same as **f** for 10-mm ball. **j**, Same as **g** for 10-mm ball.



Extended Data Figure 8 | Low rotational velocities during walking in darkness are not well captured by EBw.s activity. Comparison of angular velocity against PVA-estimated angular velocity for all flies walking in darkness on 6-mm ball (Fig. 4, see Methods). Each point is computed across a 20-frame window, and coloured based on the strength of the PVA during that epoch. Three features are prominent: (1) rotational velocity and PVA-estimated angular velocity are correlated, but with some spread and with different slopes for different flies, that is, effective walking-rotation-to-PVA gains can be different for different flies (see Extended Data Fig. 7f, i). (2) Low rotational velocities are not always well captured by EB activity which can drift under such conditions (see points near 0 of the *y* axis). (3) Most cases of EB activity drift seem to occur in phases when the PVA strength is low (as marked by dark blue points arranged in a horizontal line for low velocities).



Extended Data Figure 9 | Gain and correlation coefficients for flies walking with a bright stripe and after the stripe has disappeared. **a**, Distribution of gains between accumulated walking rotation and accumulated PVA estimate for flies walking in the dark before exposure to visual stimulus in closed-loop experiment (mean = 0.47 ± 1.2, n = 397 walking bouts). **b**, Distribution of gains between accumulated walking rotation and PVA estimate of flies walking with a bright stripe with high (light red, mean = 0.86 ± 0.64, n = 147 walking bouts) or low (light blue, mean = 0.54 ± 0.5, n = 132) closed-loop gain. All gains used were close to the likely 'natural' gain. **c**, Distribution of gains between accumulated walking rotation and PVA estimate of flies walking in darkness after walking with a stripe under closed-loop control in high (red, mean = 0.57 ± 0.84, n = 150) or low (blue, mean = 0.46 ± 0.7, n = 134) gain conditions. **d**, Distribution of correlation coefficients between accumulated walking rotation and accumulated PVA estimate for flies walking in darkness before visual experience in the closed-loop setup (mean = 0.6 ± 0.42). **e**, Distribution of correlation coefficients between accumulated walking rotation and accumulated PVA estimate for flies walking with a stripe under closed-loop control with high (light red, mean = 0.79 ± 0.34) or low (light blue, mean = 0.85 ± 0.18) closed-loop gain. **f**, Distribution of correlation coefficients between accumulated walking rotation and accumulated PVA estimate for flies walking in darkness after walking with a stripe under closed-loop control with high (red, mean = 0.48 ± 0.43) or low (blue, mean = 0.49 ± 0.49) gain. P values (Kolmogorov–Smirnov two-sample test) for tests of the null hypothesis that the correlations from two different conditions are drawn from the same distribution are as follows. The null hypothesis can be rejected at $P < 0.05$ for: $\text{gain}_{\text{DarkAfterHighGain}}$ vs $\text{gain}_{\text{DarkAfterLowGain}}$: $P = 0.04$; $\text{gain}_{\text{DarkNaive}}$ vs $\text{gain}_{\text{DarkAfterHighGain}}$: $P = 0.01$; $\text{gain}_{\text{StripeHighGain}}$ vs $\text{gain}_{\text{StripeLowGain}}$: $P = 4 \times 10^{-8}$; $\text{gain}_{\text{StripeHighGain}}$ vs $\text{gain}_{\text{DarkAfterHighGain}}$: $P = 3 \times 10^{-7}$; $\text{gain}_{\text{StripeLowGain}}$ vs $\text{gain}_{\text{DarkAfterLowGain}}$: $P = 0.05$; $\text{gain}_{\text{StripeLowGain}}$ vs $\text{gain}_{\text{DarkNaive}}$: $P = 0.001$; $\text{gain}_{\text{StripeHighGain}}$ vs $\text{gain}_{\text{DarkNaive}}$: $P = 1 \times 10^{-15}$. It cannot be rejected for: $\text{gain}_{\text{DarkNaive}}$ vs $\text{gain}_{\text{DarkAfterLowGain}}$: $P = 0.2$. Subscripts indicate conditions of the relevant experiments. DarkNaive: in darkness without previous exposure to closed-loop visual stimulus; DarkAfterLowGain: walking in darkness after a period of walking in closed loop with a single-stripe stimulus under low closed-loop gain conditions; DarkAfterHighGain: walking in darkness after a period of walking in closed loop with a single-stripe stimulus under high closed-loop gain conditions; StripeHighGain: walking with a single stripe under high closed-loop gain; StripeLowGain: walking with a single stripe under low closed-loop gain.



Extended Data Figure 10 | Maintenance of EB representation of orientation with persistent activity when the fly is standing. **a**, PVA estimate before stop compared to PVA estimate before restart for the 6-mm ball ($r = 0.5$, $P = 1 \times 10^{-29}$, $n = 449$, linear fit slope $= 0.96 \pm 0.02$, $P = 0$, intercept: 0.2 ± 0.06, $P = 0.0006$, $R^2 = 0.83$). **b**, Difference in PVA before stop and before restart plotted against duration over which the fly was standing (mean standing time, $t_{mean} = 6.6 \pm 5.1$ s, mean PVA difference, $\Delta PVA_{mean} = 0.09 \pm 1$). **c**, Same as **a** for the 10-mm ball ($r = 0.56$, $P = 1 \times 10^{-31}$, $n = 374$, intercept $= 0.1 \pm 0.06$, $P = 0.09$, slope $= 0.97 \pm 0.016$, $P = 0$, $n = 374$, $R^2 = 0.903$). **d**, Same as **b** for the 10-mm ball ($t_{mean} = 6.2 \pm 4.5$ s, $\Delta PVA_{mean} = 0.03 \pm 0.8$). **e**, PVA estimate before stop compared to PVA estimate at restart for the 10-mm ball ($r = 0.48$, $P = 1 \times 10^{-22}$, $n = 374$, slope $= 0.96 \pm 0.02$, $P = 0$, intercept $= 0.13 \pm 0.06$, $P = 0.02$, $R^2 = 0.91$). **f**, Difference in PVA estimate before stop and at restart for the 10-mm ball and duration over which the fly was standing ($t_{mean} = 6.1 \pm 4.47$ s, $\Delta PVA_{mean} = 0.04 \pm 0.9$). **g**, PVA estimate before stop compared to PVA estimate before restart during closed-loop behaviour with a single stripe ($r = 0.64$, $P = 1.5 \times 10^{-46}$, $n = 388$, intercept $= 0.03 \pm 0.07$, $P = 0.6$, slope $= 1 \pm 0.02$, $P = 0$, $R^2 = 0.85$). **h**, Difference in PVA before stop and before restart in single stripe closed-loop trial plotted against duration for which the fly was not walking ($t_{mean} = 4.85 \pm 3.0$ s, $\Delta PVA_{mean} = 0.04 \pm 0.74$). **i**, PVA estimate before stop compared to PVA estimate at restart during closed-loop behaviour with a single stripe ($r = 0.67$, $P = 5 \times 10^{-52}$, $n = 388$, intercept $= 0.1 \pm 0.06$, $P = 0.1$, slope $= 0.97 \pm 0.02$, $P = 0$, $R^2 = 0.88$). **j**, Difference in PVA estimate before stop and at restart during closed-loop behaviour with a single stripe ($t_{mean} = 4.97 \pm 3.0$ s, $\Delta PVA_{mean} = 0.02 \pm 0.65$). **k**–**n**, Same as **g**–**j** for closed-loop walking with the pattern with multiple features. **g**, $r = 0.66$, $P = 2 \times 10^{-19}$, $n = 146$, intercept $= 0.2 \pm 0.1$, $P = 0.05$, slope $= 0.9 \pm 0.03$, $P = 0$, $R^2 = 0.85$. **h**, $r = 0.6$, $P = 1.6 \times 10^{-14}$, $n = 146$, intercept $= 0.19 \pm 0.11$, $P = 0.07$, slope $= 0.91 \pm 0.03$, $P = 2.1 \times 10^{-64}$, $R^2 = 0.87$. **i**, $t_{mean} = 6.3 \pm 7.4$ s, $\Delta PVA_{mean} = -0.1 \pm 0.8$. **j**, $t_{mean} = 6.4 \pm 7.4$ s, $\Delta PVA_{mean} = -0.04 \pm 0.8$. **o**–**r**, Same as **g**–**j** for closed-loop walking with two stripes. **o**, $r = 0.6$, $P = 5.1 \times 10^{-15}$, $n = 139$, intercept $= 0.19 \pm 0.11$, $P = 0.08$, slope $= 0.93 \pm 0.03$, $P = 0$, $R^2 = 0.88$. **p**, $r = 0.7$, $P = 1.4 \times 10^{-21}$, $n = 139$, intercept $= 0.2 \pm 0.1$, $P = 0.03$, slope $= 0.95 \pm 0.03$, $P = 0$, $R^2 = 0.9$. **q**, $t_{mean} = 5.6 \pm 5.8$ s, $\Delta PVA_{mean} = 0.01 \pm 0.7$. **r**, $t_{mean} = 5.8 \pm 5.8$ s, $\Delta PVA_{mean} = 0.1 \pm 0.66$.

LETTER

doi:10.1038/nature14439

Strangulation as the primary mechanism for shutting down star formation in galaxies

Y. Peng[1,2], R. Maiolino[1,2] & R. Cochrane[1,3]

Local galaxies are broadly divided into two main classes, star-forming (gas-rich) and quiescent (passive and gas-poor). The primary mechanism responsible for quenching star formation in galaxies and transforming them into quiescent and passive systems is still unclear. Sudden removal of gas through outflows[1–6] or stripping[7–9] is one of the mechanisms often proposed. An alternative mechanism is so-called "strangulation"[10–14], in which the supply of cold gas to the galaxy is halted. Here we report an analysis of the stellar metallicity (the fraction of elements heavier than helium in stellar atmospheres) in local galaxies, from 26,000 spectra, that clearly reveals that strangulation is the primary mechanism responsible for quenching star formation, with a typical timescale of four billion years, at least for local galaxies with a stellar mass less than 10^{11} solar masses. This result is further supported independently by the stellar age difference between quiescent and star-forming galaxies, which indicates that quiescent galaxies of less than 10^{11} solar masses are on average observed four billion years after quenching due to strangulation.

Figure 1 qualitatively illustrates the expected evolution of galaxies in the two quenching scenarios. A more quantitative analysis is provided below. In the scheme of Fig. 1, at $t < t_q$ the galaxy is subject to gas inflow and forms stars, thus increasing the stellar mass and the metallicity of the gas, out of which new stars form; hence the metallicity of the newly formed stars also increases with time. The metallicity increment is modest, since inflowing gas dilutes the metal content in the interstellar medium. At $t = t_q$ quenching occurs. In the scenario of sudden gas removal (for example, expulsion of gas by a strong wind or strong ram pressure stripping) star formation is suddenly quenched (Fig. 1a), and the galaxy evolves into a quiescent system. In this case the stellar metallicity and stellar mass (M_{star}) of the quiescent galaxy are the same as those of its star-forming progenitor just before quenching. In the case of quenching by strangulation, star formation can continue, using the gas available in the galaxy until it is completely used up (Fig. 1b). During this phase the gas metallicity increases more steeply than in the previous case because of the lack of dilution from inflowing gas. The stellar mass also increases slightly. The general product of strangulation is a quiescent galaxy with a stellar metallicity that is much higher than its star-forming progenitor, and a slightly higher stellar mass.

Observationally, obviously we cannot follow the evolution of the stellar metallicity of individual galaxies. However, we can statistically investigate the metallicity difference between star-forming and quiescent galaxies. This statistical approach has already been successfully exploited, for instance, to investigate the dependence of the quenching mechanism on mass and environment[15]. We have used the Sloan Digital Sky Survey (SDSS) spectra of local (redshift $z < \sim 0.1$) galaxies to extract a subsample of 3,905 star-forming and 22,618 quiescent galaxies whose spectra have a signal to noise ratio of $S/N > 20$ per spectral pixel, which ensures a reliable determination of the stellar metallicities (details of the sample selection and determination of the stellar metallicities are given in the Methods).

Figure 2a shows the average stellar metallicity of star-forming (blue line) and quiescent (red line) galaxies as a function of stellar mass, obtained by using a sliding average of 0.2 dex in M_{star} (error bars give the 1σ uncertainty of the mean stellar metallicity). At a given stellar mass, the stellar metallicity of quiescent galaxies is noticeably higher than for star-forming galaxies, at least for $M_{star} < 10^{11} M_{\odot}$, where $M_{\odot}$ is the solar mass. This is not what is expected in the case of sudden gas removal, but it is qualitatively consistent with the strangulation scenario. Below we investigate more quantitatively the agreement of the data with the strangulation scenario.

For a system forming stars without any inflow, the temporal evolution of the gaseous and stellar metallicity, as well as of the stellar mass, can be trivially solved analytically, as discussed in the Methods. The key parameters are: (1) the total gas mass at the time of quenching, $M_{gas}(t_q)$ or, equivalently, the gas fraction $f_{gas}(t_q)$; (2) the global efficiency of star formation ε, defined as the star-formation rate (SFR) $= \varepsilon M_{gas}$ (where M_{gas} is in this case the total gas mass, both atomic and molecular); (3) the amount of any outflowing gas that is lost (that is, which does not fall back onto the galaxy and is not recycled), which can be approximated as being proportional to the SFR and parameterized through the so-called outflow mass loading factor λ, defined as $M_{outflow} = \lambda$SFR (note that if part of the gas falls back and is recycled[16,17], then λ would account only for the fraction of gas that is lost, that is, it is an 'effective' outflow loading factor).

We are dealing with differential quantities (in particular, differential stellar metallicities ΔZ_{star}), so the metallicity of the gas and stars at the beginning of the quenching ($t = t_q$) is unimportant. Yet, the value of M_{star} at $t = t_q$ is relevant because it defines the gas fraction, and hence the gas mass available for further star formation. Indeed, it is observationally well known that the gas fraction in star-forming galaxies decreases with M_{star} (refs 18–21; see Methods for the detailed functional form). The global star-formation efficiency ε is also known from observations (see Methods).

Figure 2b shows the stellar metallicity difference, as a function of stellar mass (solid thin lines in colours), as expected from the strangulation scenario, at five different times after the quenching/strangulation time ($\Delta t = t - t_q$). In this plot we are assuming that the effective outflow loading factor is $\lambda = 0$ (that is, any outflowing gas falls back and is recycled), implying that the system behaves as a closed box after strangulation (no inflow and no effective outflow). The case of substantial outflow after strangulation is discussed below. The decline of the curves at high M_{star} is primarily a consequence of the gas fraction of star-forming galaxies decreasing as a function of M_{star}: massive galaxies have low gas content, and so once such a galaxy has been strangled, its available gas can produce few stars relative to those already present, and the average stellar metallicity is not much affected. The thick black line with error bars shows the observed stellar metallicity difference between quiescent and star-forming galaxies (that is, the difference between red and blue data in Fig. 2a). The observed difference is consistent, within uncertainties, with the strangulation scenario in which quiescent galaxies at $M_{star} < 10^{11} M_{\odot}$ are,

[1]Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Avenue, Cambridge CB3 0HE, UK. [2]Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. [3]Institute for Astronomy, Royal Observatory Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK.

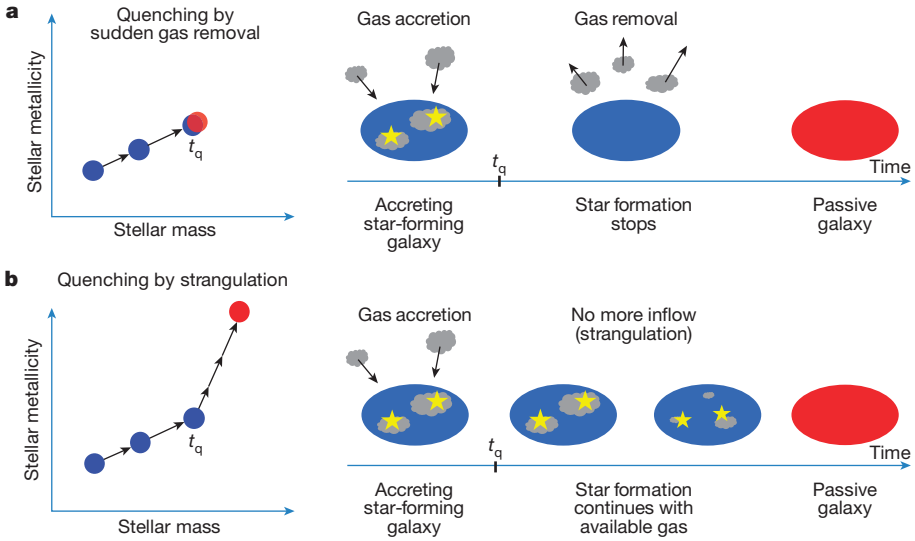


Figure 1 | Illustration of two different quenching scenarios and their effect on stellar metallicities. **a**, Rapid and complete removal of the gas reservoir of the galaxy (for example, from strong outflows or ram pressure stripping) results in a passive galaxy with the same mass and same stellar metallicity as its star-forming progenitor. **b**, In the strangulation scenario, the galaxy can continue to form stars from the available enriched interstellar medium and, as a consequence, increase its stellar mass and stellar metallicity.

on average, observed 4 billion years (Gyr) after quenching due to strangulation.

This result can be tested independently by looking at the stellar age difference between quiescent and star-forming galaxies, which can be measured from the galaxy spectra, independently of the stellar metallicities (see Methods). Figure 3a shows the average age of all galaxies (black line), as a function of stellar mass. The average age increases steeply with M_{star}, but this is mostly because the fraction of quiescent-old galaxies, relative to star-forming young galaxies, increases with stellar mass. If the population of galaxies is split into quiescent and star-forming samples, then the average ages of the two samples show a much flatter dependence on M_{star}, as illustrated by the red (quiescent) and blue (star-forming) lines in Fig. 3a. Most importantly, as illustrated in Fig. 3b, the age difference between the two populations is reasonably constant and equal to about 4 Gyr. The latter is the same age difference required to explain the stellar metallicity difference in the strangulation scenario; it therefore further and independently confirms this scenario. The relatively long timescale of about 4 Gyr is also supported by independent observations and simulations[22–25].

We also investigate the case of a substantial 'effective' outflow, after strangulation, by setting $\lambda = 1$, which is a typical loading factor observed in star-forming galaxies[26–28]. Figure 4 shows the resulting ΔZ_{star} curves, which are completely inconsistent with the observed data. This further confirms that gas removal by outflows plays a minor part in quenching galaxies. This includes any external environmental effect such as gas removal in satellite galaxies when falling into a more massive halo, or feedback process such as outflow driven by an active galactic nucleus.

Overall, our results strongly support the scenario in which local quiescent galaxies with $M_{star} < 10^{11} M_{\odot}$ (that is, the vast majority of galaxies) are primarily quenched as a consequence of strangulation. However, this analysis does not clarify what the strangulation mechanism is (for example, hot halo environmental strangulation or strangulation via various preventive feedback mechanisms, such as circumgalactic gas heating[29]). Additional analysis is needed to investigate the strangulation mechanism (for example, by studying the central/satellite and environmental dependence; see Methods subsection 'Stellar metallicity for central and satellite galaxies'). We also note that

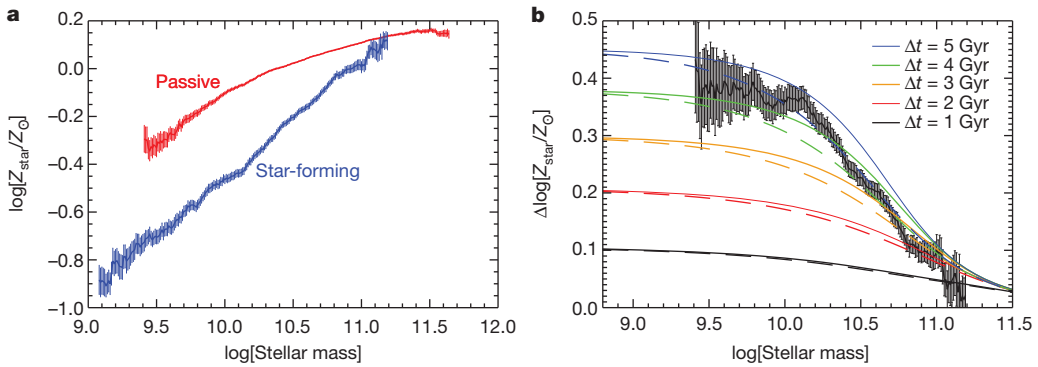


Figure 2 | Stellar metallicities for star-forming and quiescent galaxies. **a**, Average stellar metallicity as a function of stellar mass for all star-forming galaxies (thick blue line with error bars) and all quiescent galaxies (thick red line with error bars) for galaxies at $<z> \approx 0.05$. Error bars correspond to the 1σ error on the mean value. **b**, Average metallicity difference between all star-forming and all quiescent galaxies (thick black line with error bars). Error bars on the black line indicate the 1σ uncertainty in the metallicity difference. The metallicity difference decreases with increasing stellar mass. It reaches the maximum value around 0.4 dex for galaxies at $M_{star} \approx 10^{9.5} M_{\odot}$ and becomes negligible at $M_{star} \geq 10^{11} M_{\odot}$. The coloured lines show the metallicity difference predicted by a simple close-box model at different times Δt after strangulation. Solid lines are for the final mass (at time $t = t_q + \Delta t$), while dashed lines are for the mass at strangulation ($t = t_q$). The observed mass-dependent metallicity difference between quiescent and star-forming galaxies (thick black line) can be very well reproduced by a close-box model with a constant $\Delta t \approx 4$ Gyr, largely independent of stellar mass.

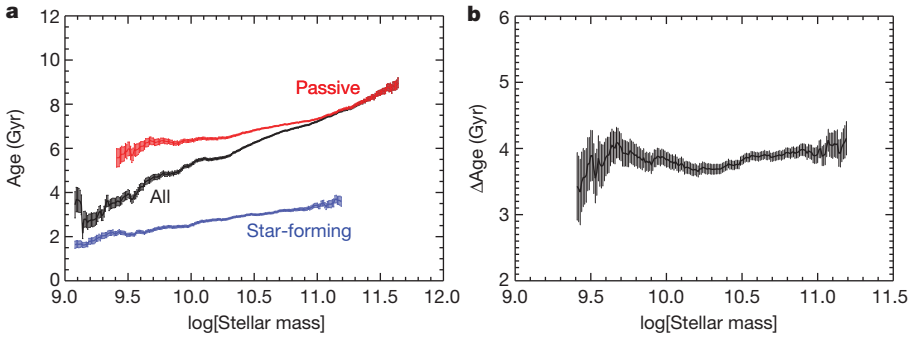


Figure 3 | Stellar ages for star-forming and quiescent galaxies. **a**, Luminosity-weighted stellar age as a function of stellar mass for star-forming (blue line), quiescent (red line) and all galaxies (black line) at $<z> \approx 0.05$. Error bars are the 1σ error on the mean value. The average age for all galaxies (black line) strongly depends on the stellar mass, largely due to the fact that the red fraction (that is, the mixture of quiescent and star-forming galaxies) strongly depends on stellar mass. However, the dependence on stellar mass becomes much weaker once the whole sample is split into star-forming and quiescent galaxies. **b**, Average age difference between quiescent and star-forming galaxies as a function of stellar mass. Error bars on the black line indicate the 1σ uncertainty in the age difference. Remarkably, the age difference for all galaxies is largely independent of mass, with a mean value of around 4 Gyr, which is consistent with the mass-independent time Δt from strangulation required to explain the difference in stellar metallicities.

our results do not imply any claim about the morphological changes of the galaxy population, as it is not completely clear whether the morphological transformation is associated with star-formation quenching[30].

The data presented in this paper cannot shed light on the quenching mechanism at $M_{\rm star} \geq 10^{11} M_{\odot}$. At $M_{\rm star} \approx 10^{11} M_{\odot}$ the stellar metallicity of quenched galaxies is similar to the stellar metallicity of star-forming galaxies (Fig. 2), which can be interpreted equally well as quenching by sudden gas removal (such as by outflows; Fig. 1a) or as quenching by strangulation of gas-poor massive galaxies (indeed, in massive galaxies the small amount of available gas, as shown in Extended Data Fig. 1a, does not allow much star formation, and hence there is little variation of the stellar metallicity, even if the galaxy is strangled; see also discussions in Methods subsection 'Fraction of galaxies quenched by rapid gas removal'). At even higher stellar masses our analysis is not feasible, since star-forming galaxies with $M_{\rm star} > 10^{11} M_{\odot}$ are extremely rare in the local Universe, hence preventing our statistical approach. To shed light on the quenching mechanism of massive galaxies a similar analysis has to be performed at high z, where massive star-forming galaxies are abundant and gas-rich.

The results obtained in this paper apply on average to the bulk of the local galaxy population. However, for individual galaxies, other quenching mechanisms such as fast gas removal via outflows (which can also help to explain the α-element enhancement in massive elliptical galaxies) and environmental effects (such as ram-pressure stripping, tidal stripping, harassment and mergers), may work together with, or cause, strangulation to shape the detailed quenching process (see Methods subsection 'Fraction of galaxies quenched by rapid gas removal'). These additional quenching mechanisms may modify the amount of stellar metallicity enhancement, the quenching timescale, and/or the gas content, which may all contribute to the scatter in the stellar metallicity and age differences between star-forming and quiescent galaxies.

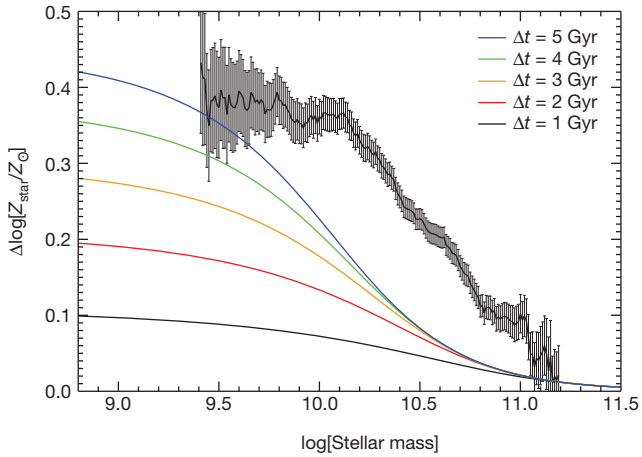


Figure 4 | The effect of outflows on stellar metallicity evolution. As for Fig. 2b, but for the case of an effective mass-loading factor $\lambda = 1$ after strangulation. In this case, the observed stellar metallicity difference (black thick line) is clearly not reproduced by the model, further suggesting that outflows do not play a major part. Error bars on the black line indicate the 1σ uncertainty in the metallicity difference.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 November 2014; accepted 16 March 2015.

1. Di Matteo, T., Springel, V. & Hernquist, L. Energy input from quasars regulates the growth and activity of black holes and their host galaxies. *Nature* **433,** 604–607 (2005).
2. Hopkins, P. F. *et al.* Unified, merger-driven model of the origin of starbursts, quasars, the cosmic X-ray background, supermassive black holes, and galaxy spheroids. *Astrophys. J. Suppl. Ser.* **163,** 1–49 (2006).
3. Maiolino, R. *et al.* Evidence of strong quasar feedback in the early Universe. *Mon. Not. R. Astron. Soc.* **425,** L66–L70 (2012).
4. Diamond-Stanic, A. *et al.* High-velocity outflows without AGN feedback: Eddington-limited star formation in compact massive galaxies. *Astrophys. J.* **755,** L26 (2012).
5. Förster Schreiber, N. M. *et al.* The Sins/zC-Sinf survey of $z \sim 2$ galaxy kinematics: evidence for powerful active galactic nucleus-driven nuclear outflows in massive star-forming galaxies. *Astrophys. J.* **787,** 38 (2014).
6. Cicone, C. *et al.* Massive molecular outflows and evidence for AGN feedback from CO observations. *Astron. Astrophys.* **562,** A21 (2014).
7. Gunn, J. E. & Gott, J. R. On the infall of matter into clusters of galaxies and some effects on their evolution. *Astrophys. J.* **176,** 1–19 (1972).
8. Abadi, M. G., Moore, B. & Bower, R. G. Ram pressure stripping of spiral galaxies in clusters. *Mon. Not. R. Astron. Soc.* **308,** 947–954 (1999).
9. Quilis, V., Moore, B. & Bower, R. Gone with the wind: the origin of S0 galaxies in clusters. *Science* **288,** 1617–1620 (2000).
10. Larson, R. B., Tinsley, B. M. & Caldwell, C. N. The evolution of disk galaxies and the origin of S0 galaxies. *Astrophys. J.* **237,** 692–707 (1980).
11. Balogh, M. L. & Morris, S. L. H. α photometry of Abell 2390. *Mon. Not. R. Astron. Soc.* **318,** 703–714 (2000).
12. Balogh, M. L., Navarro, J. F. & Morris, S. L. The origin of star formation gradients in rich galaxy clusters. *Astrophys. J.* **540,** 113–121 (2000).
13. Kereš, D., Katz, N., Weinberg, D. H. & Davé, R. How do galaxies get their gas? *Mon. Not. R. Astron. Soc.* **363,** 2–28 (2005).
14. Dekel, A. & Birnboim, Y. Galaxy bimodality due to cold flows and shock heating. *Mon. Not. R. Astron. Soc.* **368,** 2–20 (2006).
15. Peng, Y. *et al.* Mass and environment as drivers of galaxy evolution in SDSS and zCOSMOS and the origin of the Schechter function. *Astrophys. J.* **721,** 193–221 (2010).
16. De Lucia, G., Kauffmann, G. & White, S. D. M. Chemical enrichment of the intracluster and intergalactic medium in a hierarchical galaxy formation model. *Mon. Not. R. Astron. Soc.* **349,** 1101–1116 (2004).

17. Oppenheimer, B. D. *et al.* Feedback and recycled wind accretion: assembling the $z = 0$ galaxy mass function. *Mon. Not. R. Astron. Soc.* **406,** 2325–2338 (2010).
18. Baldry, I. K., Glazebrook, K. & Driver, S. P. On the galaxy stellar mass function, the mass-metallicity relation and the implied baryonic mass function. *Mon. Not. R. Astron. Soc.* **388,** 945–959 (2008).
19. Peeples, M. S. & Shankar, F. Constraints on star formation driven galaxy winds from the mass-metallicity relation at $z = 0$. *Mon. Not. R. Astron. Soc.* **417,** 2962–2981 (2011).
20. Boselli, A. *et al.* Cold gas properties of the Herschel Reference Survey. II. Molecular and total gas scaling relations. *Astron. Astrophys.* **564,** A66 (2014).
21. Santini, P. *et al.* The evolution of the dust and gas content in galaxies. *Astron. Astrophys.* **562,** A30 (2014).
22. Wetzel, A. R., Tinker, J. L., Conroy, C. & van den Bosch, F. C. Galaxy evolution in groups and clusters: satellite star formation histories and quenching time-scales in a hierarchical Universe. *Mon. Not. R. Astron. Soc.* **432,** 336–358 (2013).
23. Hirschmann, M. *et al.* The influence of the environmental history on quenching star formation in a Λ cold dark matter universe. *Mon. Not. R. Astron. Soc.* **444,** 2938–2959 (2014).
24. Schawinski, K. *et al.* The green valley is a red herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies. *Mon. Not. R. Astron. Soc.* **440,** 889–907 (2014).
25. Woo, J., Dekel, A., Faber, S. M. & Koo, D. C. Two conditions for galaxy quenching: compact centres and massive haloes. *Mon. Not. R. Astron. Soc.* **448,** 237–251 (2015).
26. Davé, R., Finlator, K. & Oppenheimer, B. D. Galaxy evolution in cosmological simulations with outflows—II. Metallicities and gas fractions. *Mon. Not. R. Astron. Soc.* **416,** 1354–1376 (2011).
27. Hopkins, P. F., Quataert, E. & Murray, N. Stellar feedback in galaxies and the origin of galaxy-scale winds. *Mon. Not. R. Astron. Soc.* **421,** 3522–3537 (2012).
28. Cicone, C. *et al.* Massive molecular outflows and evidence for AGN feedback from CO observations. *Astron. Astrophys.* **562,** A21 (2014).
29. Davé, R., Finlator, K. & Oppenheimer, B. D. An analytic model for the evolution of the stellar, gas and metal content of galaxies. *Mon. Not. R. Astron. Soc.* **421,** 98–107 (2012).
30. Carollo, C. M. *et al.* ZENS IV. Similar morphological changes associated with mass- and environment-quenching, and the relative importance of bulge growth versus the fading of disks. *Astrophys. J.* (submitted).

Acknowledgements We thank A. Gallazzi and her collaborators for making their SDSS DR4 version of the stellar ages and metallicities catalogues publicly available. We thank S. Lilly, A. Renzini, H.-W. Rix and M. Haehnelt for useful discussions. We acknowledge NASA's IDL Astronomy Users Library, the IDL code base maintained by D. Schlegel, and the *k*correct software package of M. Blanton.

Author Contributions Y.P. and R.M. co-developed the idea; both contributed to the interpretation and manuscript writing. Y.P. and R.C. contributed to the measurements and analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.P. (yp244@mrao.cam.ac.uk).

METHODS

Sample and observational data. The parent sample of galaxies analysed in this paper is the SDSS DR7 sample that is used in refs 15 and 31 for similar statistical investigations of the quenching process. Briefly, it is a magnitude-selected sample of galaxies in the redshift range of $0.02 < z < 0.085$ that have clean photometry and Petrosian SDSS r-band magnitudes[41] in the range $10.0 < r < 18.0$ after correcting for Galactic extinction. The parent photometric sample contains 1,579,314 objects after removing duplicates, of which 238,474 have reliable spectroscopic redshift measurements. As a consequence of the relatively broad redshift range $0.02 < z < 0.085$, the projected physical aperture of the SDSS spectroscopic fibre changes substantially across the sample. To test whether there are any noticeable aperture effects, we further split the whole redshift range into two narrower redshift ranges, at $0.02 < z < 0.05$ and $0.05 < z < 0.085$. The results are shown in the Extended Data Fig. 2. It is clear that the results change little as a function of redshift.

Each galaxy is weighted by $1/\mathrm{TSR} \times 1/V_{\max}$, where TSR is a spatial target sampling rate, determined using the fraction of objects that have spectra in the parent photometric sample within the minimum SDSS fibre spacing of 55 arcsec of a given object. The $V_{\max}$ values are derived from the k-correction program v4_1_4 (ref. 32). The use of $V_{\max}$ weighting allows us safely to include representatives of the galaxy population down to a stellar mass of about $10^9 M_{\odot}$.

The stellar masses are determined from the k-correction code using population synthesis models[33] and a Chabrier initial mass function. The galaxy population is then divided into star-forming and passive galaxies, based on their spectroscopic emission line classifications and rest-frame $(U - B)$ colours. Star-forming galaxies are defined to have the classification flag (that is, the "iclass" keyword), in the SFR catalogue[34], set to 1. This also excludes those objects hosting an active galactic nucleus, for which the derived metallicities are probably not reliable. Passive galaxies are defined to have their $(U-B)$ colours redder than a threshold colour that is given by equation (2) in ref. 15 and without Hα emission (that is, undetected with $S/N < 3$).

The stellar metallicities and r-band weighted stellar ages were derived by ref. 35 from the spectral absorption features of SDSS Data Release 4 spectra, which are then cross-matched with our parent galaxy sample. We refer to ref. 35 for a detailed description of the method. Here, we mention only that the method consists of measuring the strength of a set of carefully selected spectral absorption features, including several well calibrated Lick indices and the 4,000 Å break. Then they[35] compute the median-likelihood estimates of the stellar metallicities and r-band light weighted ages by comparing the spectral absorption features to a large library of 150,000 Monte Carlo realizations spanning a full range of physically plausible star-formation histories. The stellar metallicities are derived for both passive and star-forming galaxies, for which the contamination of stellar absorption features by nebular emission has been carefully removed. To ensure a reliable metallicity measurement, we restrict our galaxy sample to galaxies that have a median S/N per pixel of at least 20 over the whole spectrum. This ensures the average uncertainty of the stellar metallicity and light-weighted age to be less than ±0.15 dex. The requirement on the S/N could, in principle, be lowered to increase the statistics, but this would result in higher uncertainties in the metallicity and age measurements. Since the statistical errors on the mean metallicity and mean age are already reasonably small with the above S/N selection, the analysis benefits more from solid measurements than from improved statistics.

With the selection criteria given above, the final sample consists of 22,618 passive galaxies and 3,905 star-forming galaxies. All of these galaxies have reliable stellar mass, stellar metallicities and age measurements.

Metallicity evolution during quenching via strangulation. The quenching process via strangulation can be quantitatively described by using the analytical framework discussed in refs 36 and 37. These simple analytical models take into account the key physical processes of inflow, star formation, outflow and metal production.

We assume the star-formation law holds in the same way before and during the quenching process, that is, the instantaneous average SFR of the galaxy is always related to the gas mass present within the galaxy as:

$$\mathrm{SFR} = \varepsilon M_{\mathrm{gas}} \tag{1}$$

where M_{gas} is the total gas mass (both atomic and molecular). In fact, equation (1) can be regarded as the definition of effective, global star-formation efficiency ε or, more properly, the inverse of the total gas depletion timescale $\tau_{\mathrm{dep}} = M_{\mathrm{gas}}/\mathrm{SFR} = 1/\varepsilon$. Note that, from equation (1), the specific SFR can be expressed as: specific SFR = $\mathrm{SFR}/M_{\mathrm{star}} = \varepsilon\, M_{\mathrm{gas}}/M_{\mathrm{star}}$.

The mass-loss rate of the galaxy Ψ, that is, the outflow rate, is very likely to be closely related to the average SFR of the galaxy. Analogously to equation (1), we link these two quantities together via λ, to give:

$$\Psi = \lambda \mathrm{SFR} \tag{2}$$

where λ is the mass-loading factor. Similar to ε, equation (2) can be regarded as the definition of λ. It should be noted that strangulating the gas inflow does not necessarily turn the gas regulator model to a simple close-box model, since the galaxy can continue to have an outflow.

The general evolution of the gas metallicity Z_{gas}, without assuming any equilibrium condition, is given by equation (32) in ref. 37, that is:

$$\frac{\mathrm{d}Z_{\mathrm{gas}}}{\mathrm{d}t} = y\varepsilon - (Z_{\mathrm{gas}} - Z_0)\frac{\Phi}{M_{\mathrm{gas}}} \tag{3}$$

where y is the average yield per stellar generation and is assumed to be a constant, Φ is the inflow rate, Z_0 is the metallicity of the infalling gas and M_{gas} is the gas mass. When quenching via strangulation starts, Φ is set to zero and equation (3) can be easily solved:

$$Z_{\mathrm{gas}}(t) = Z_{\mathrm{gas}}(t_q) + y\varepsilon t \tag{4}$$

where $Z_{\mathrm{gas}}(t_q)$ is the gas metallicity at the time when quenching begins. It is clear from equation (4) that $\Delta Z_{\mathrm{gas}} = y\varepsilon t$, that is, for a constant yield and star-formation efficiency, the gas metallicity increase is simply proportional to time. We note that when the inflow is truncated, the gas metallicity is independent of the outflow.

From equation (4), the logarithmic increase of the gas metallicity when quenching begins is given by:

$$\log Z_{\mathrm{gas}}(t) - \log Z_{\mathrm{gas}}(t_q) = \log\left[1 + \frac{y\varepsilon t}{Z_{\mathrm{gas}}(t_q)}\right] \tag{5}$$

Before the start of the quenching, according to equation (35) in ref. 37, Z_{gas} is proportional to the yield y if $Z_0 \approx 0$. By inserting it into the right-hand side of equation (5), y will cancel out from both the numerator and denominator. Therefore, when the inflow is truncated, the amount of logarithmic increase of the gas metallicity is independent of the yield. Hence, any uncertainty on the yield is completely irrelevant to our analysis.

The general evolution of the stellar metallicity Z_{star}, without assuming any equilibrium condition, is given by equation (40) in ref. 37 as:

$$\begin{aligned}\frac{\mathrm{d}Z_{\mathrm{star}}}{\mathrm{d}t} &= \mathrm{sSFR} \times (1-R)(Z_{\mathrm{gas}} - Z_{\mathrm{star}}) \\ &= \frac{\varepsilon M_{\mathrm{gas}}}{M_{\mathrm{star}}}(1-R)(Z_{\mathrm{gas}} - Z_{\mathrm{star}})\end{aligned} \tag{6}$$

where R is the fraction of the mass of the newly formed stars that is quickly returned to the interstellar medium through stellar winds and supernovae. It is clear from equation (6) that the stellar metallicity simply evolves towards the gas metallicity on a timescale controlled by $(\mathrm{specific\ SFR})^{-1}$.

The change of stellar mass of the galaxy per unit time is given by:

$$\frac{\mathrm{d}M_{\mathrm{star}}}{\mathrm{d}t} = (1-R)\mathrm{SFR}$$

where $(1-R)$SFR is the net SFR that contributes to the net stellar mass increase of the galaxy, that is, the fraction of newly produced stars in the form of long-lived stars. The change of the gas mass of the galaxy per unit time is given by:

$$\begin{aligned}\frac{\mathrm{d}M_{\mathrm{gas}}}{\mathrm{d}t} &= -(1-R) \times \mathrm{SFR} - \Psi \\ &= -(1-R+\lambda)\varepsilon M_{\mathrm{gas}}\end{aligned}$$

To calculate the change of the stellar metallicity during quenching, we need to know, at a given stellar mass, the gas mass (or equivalently the gas fraction), the star-formation efficiency ε and the mass-loading factor λ.

For λ, we first assumed that, during the strangulation process, the galaxy can recycle all the outflow gas by setting $\lambda = 0$, that is, we assumed a close-box model. Then, as discussed in the text, we also investigated the case of $\lambda = 1$ during strangulation.

Both gas fraction and star-formation efficiency ε (or equivalently the gas depletion timescale τ_{dep}) have been measured observationally[18–21]. In fact, the predicted gas fraction and ε determined using the model in ref. 37 match the latest observations in the local Universe[20] extremely well, as shown in the Extended Data Fig. 1. We stress again that the gas fraction and star-formation efficiency in all our calculations are defined with the total gas mass (including both atomic and molecular). While in some previous work, such as ref. 38, the average gas depletion time is found to be constant at about 2 Gyr, which refers to the molecular gas depletion time (see subsection 'Effect of constant ε on stellar metallicity evolution' below).

The change in stellar metallicity as a function of stellar mass at different times Δt after strangulation is shown in Fig. 2b (for $\lambda = 0$). During strangulation galaxies will continue to form stars with the available gas following the star-formation law given by equation (1) and their stellar mass will hence continue to grow. The coloured solid lines show the stellar metallicity increase as a function of the final

stellar mass, while the coloured dashed lines show the stellar metallicity increase as a function of the stellar mass at the epoch when the strangulation starts. As shown in Fig. 2b, the stellar metallicity increase is slightly larger if the stellar mass considered is the final stellar mass, but there are no substantial differences between these two stellar masses.

At a given stellar mass, the amount of stellar metallicity increase is nearly proportional to the time Δt elapsed from the beginning of strangulation. At a given Δt the stellar metallicity increase is larger for low-mass galaxies than for more-massive galaxies. This is because the variation of stellar metallicity depends on the size of the gas reservoir, that is, the gas fraction, at the epoch when the strangulation starts (the larger the gas reservoir, the more metals can be produced in the strangulation phase). For massive galaxies, which have a low gas fraction (Extended Data Fig. 1a), the relative amount of gas available for star formation is small, while the existing stellar population (with relatively low metallicity) is large compared to the amount of new stars (with higher metallicity) that will form during the strangulation. Therefore, although the star-formation efficiency is higher for massive galaxies (Extended Data Fig. 1b), the increase in stellar metallicity is smaller for massive galaxies than for low-mass galaxies (Fig. 2b).

It is evident, as discussed in the main text, that the observed metallicity difference can be very well reproduced by a simple close-box model with a constant mass-independent $\Delta t \approx 4$ Gyr across the entire observed range of stellar masses, which is consistent with the age difference between the two populations (Fig. 3b). We stress that a different stellar metallicity calibration and age calibration method may give different scales and different slopes of the stellar mass versus metallicity/age relations, but the results illustrated above are preserved regardless of the adopted calibration. This is because our results are mainly based on metallicity differences and age differences between star-forming and passive galaxies, that is, we are dealing with differential quantities, and therefore uncertainties in the metallicity/age scale are much less critical than in studies dealing with absolute quantities.

Finally, we discuss one second-order effect that we have not taken into consideration, but which would further reinforce our results. The star-forming progenitors of passive (quenched) galaxies observed locally should be star-forming galaxies at $z \approx 0.5$ (that is, 4 Gyr ago), and not the star-forming population observed locally in SDSS. Unfortunately, SDSS does not have the sensitivity to deliver star-forming galaxies at high redshifts in large numbers and with the same S/N as local star-forming galaxies. However, star-forming galaxies at $z \approx 0.5$ should have a metallicity even lower than local galaxies. Therefore, if any, the metallicity difference between passive and star-forming galaxies observed in Fig. 2 should be even larger. However, this effect is expected to be very small, since the mass–metallicity relation evolves very little from $z = 0$ to $z = 0.5$ (refs 39 and 40).

Stellar metallicity for central and satellite galaxies. Since a distinction between central and satellite galaxies appears in many theoretical models for the evolution of galaxies and may potentially shed light on the strangulation mechanism, we further divide the whole sample into central and satellite galaxies and the results are shown in the Extended Data Fig. 3. As discussed in ref. 31, there are difficulties in the identification of true central galaxies due to over-fragmentation of groups by the group-finding algorithm, which would lead to some satellites being misidentified as central galaxies. This effect is expected to be most severe for low-mass galaxies in high-density regions. To obtain a clean sample of true central galaxies, we further select the central galaxies in the fields, where their over-densities are below the mean over-density of the local Universe.

The stellar metallicity enhancement of satellites is slightly larger than that of central galaxies at $M_{\rm star} < 10^{10} M_{\odot}$. This suggests that low-mass satellites are likely to suffer more from the strangulation than central galaxies and hence implies an environmental origin of strangulation at these low masses (for example, inflow of gas being halted as satellite galaxies plunge into the hot halos). At masses $M_{\rm star} > 10^{10} M_{\odot}$ no difference between central galaxies and satellites is detected. This suggests that at high masses the strangulation process may operate similarly for both central galaxies and satellites. A more detailed analysis, in different environments, is required to unveil the origin of strangulation at high masses, which will be presented in a future work.

Fraction of galaxies quenched by rapid gas removal. We argue here that the majority of local quiescent galaxies with $M_{\rm star} < 10^{11} M_{\odot}$ is primarily quenched as a consequence of strangulation. This is a statistical statement, based on the average properties of the galaxy population. It is, however, interesting to quantify the fraction of galaxies whose data may potentially allow for rapid quenching by sudden gas removal (which does not cause a metallicity change) as an alternative viable process.

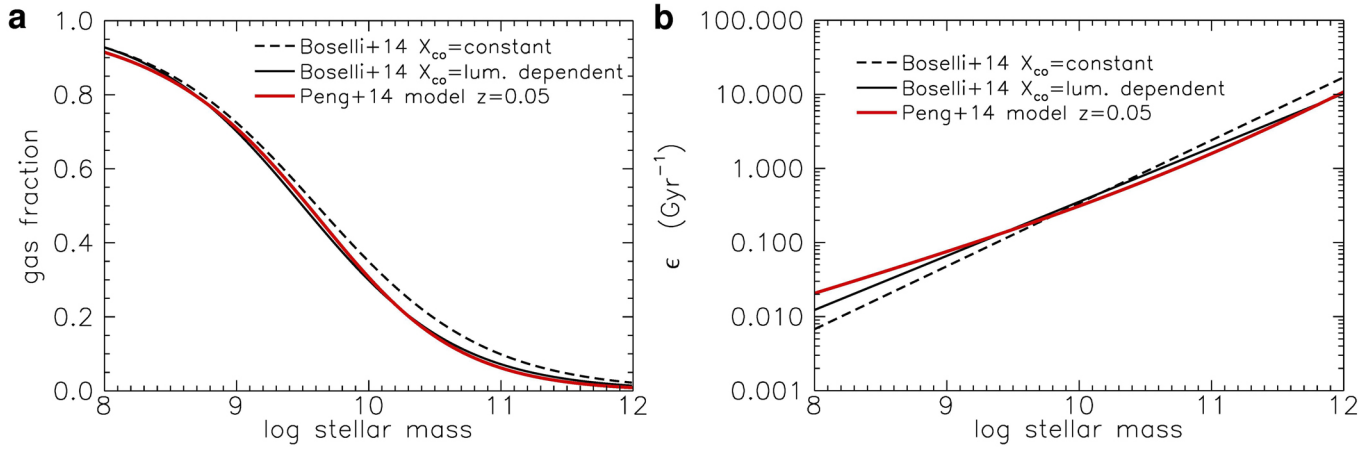
In each panel of Extended Data Fig. 4, we show the probability density function (PDF) of the stellar metallicity of star-forming galaxies (blue line) and that of the passive galaxies (red line) at a given stellar mass. Each PDF is normalized (that is, the area beneath each curve is unity). The overlapping region of the two PDFs is shaded (light red). The fraction of the shaded over the total area given by the blue PDF is noted as $f_{\rm max}$ in the label. If all star-forming galaxies were eventually to be quenched, the star-forming PDF (blue line) should eventually evolve into the passive PDF (red line).

The evident difference between star-forming and passive metallicity implies that strangulation is the primary quenching process. The opposite is not necessarily true: a similar metallicity for star-forming and passive galaxies may be caused either by sudden gas removal (such as outflows and gas stripping) or strangulation of galaxies with modest gas content. Furthermore, we also note that if the metallicity of a star-forming galaxy is similar to those of passive ones, this could mean that this galaxy initially had large gas content, was strangled, and we are observing it during the last stage of star formation in its strangulation phase. Therefore the shaded area, and hence $f_{\rm max}$, give a very conservative upper limit to the fraction of galaxies for which sudden gas removal can potentially be an alternative quenching mechanism.

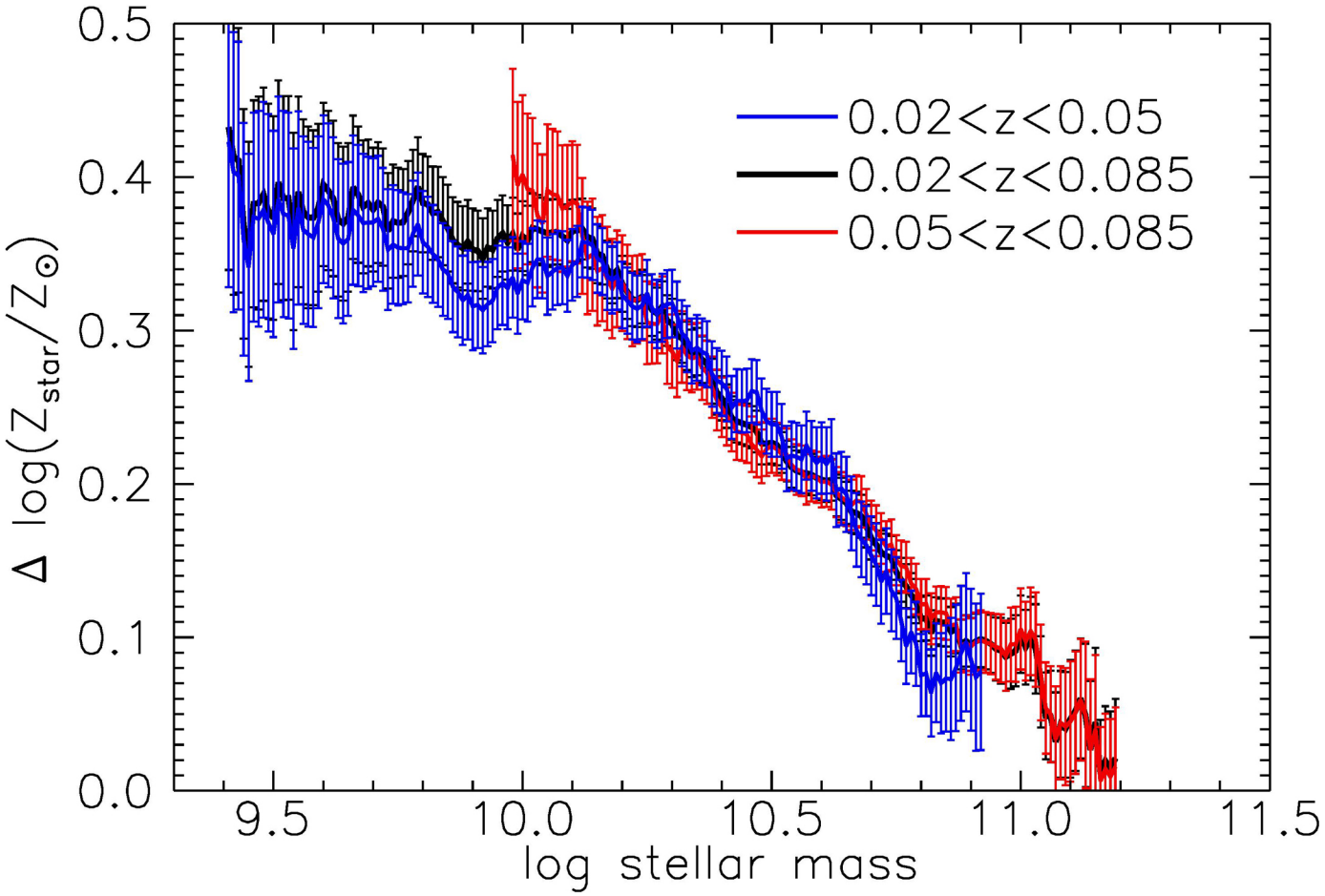
$f_{\rm max}$ is 50% at low masses, confirming that most of the galaxies at low masses must be quenched by strangulation. We recall, as discussed above, that this is a very conservative approach. $f_{\rm max}$ progressively increases with increasing stellar mass. This effect is due to the fact that the observed stellar metallicity enhancement decreases with increasing stellar mass, as shown in Fig. 2b, in which the lines with different Δt converge to zero with increasing mass. This implies that, as already discussed, the constraining power of the stellar metallicity data on the quenching mechanism decreases with increasing mass. In the most-massive cases of $M_{\rm star} > 10^{11} M_{\odot}$, the stellar metallicity data cannot shed light on the quenching mechanism, because the small metallicity difference can be interpreted equally well as rapid quenching by sudden gas removal or as quenching by strangulation.

Effect of constant ε on stellar metallicity evolution. As discussed in the subsection 'metallicity evolution during quenching via strangulation', although the star-formation efficiency ε defined with the total gas mass ($\varepsilon = {\rm SFR}/(M_{\rm HI} + M_{\rm H2})$) is very unlikely to be constant, it is useful to test the validity of the result further in Fig. 2b against a putative constant ε. We keep all other parameters unchanged in the model, except that when strangulation starts we set ε to be constant at 0.5 Gyr^{-1}, that is, a constant gas depletion timescale of $\tau_{\rm dep} = 2$ Gyr. The results are shown in the Extended Data Fig. 5. It is clear that at $M_{\rm star} \gtrsim 10^{10} M_{\odot}$, the prediction from the model is very similar to Fig. 2b and is still consistent with $\Delta t \approx 4$ Gyr or 5 Gyr. This is because the average ε in the Extended Data Fig. 1b at $M_{\rm star} \gtrsim 10^{10} M_{\odot}$ is roughly 0.5 Gyr^{-1}. However, below about $10^{10} M_{\odot}$, the model now predicts a much steeper metallicity increase, owing to the adopted constant value of $\varepsilon = 0.5$ Gyr^{-1} that is much higher than the one shown in the Extended Data Fig. 1b. This suggests that, to achieve the same observed metallicity enhancement, a higher ε will need a shorter Δt. In other words, the strangulation process can be fast (for example, less than 1 Gyr) if the ε is large (for instance, at higher redshifts[21]).

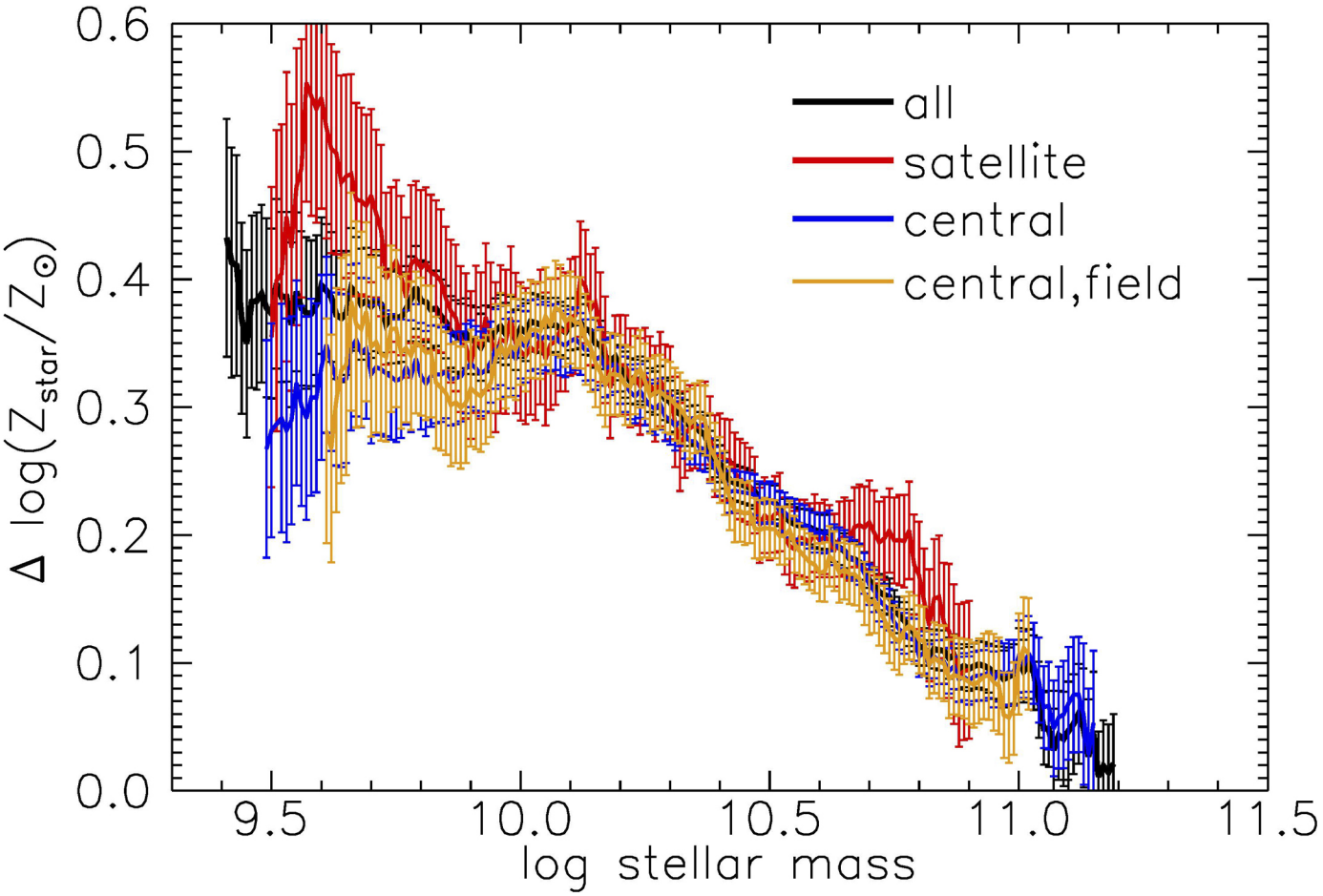
31. Peng, Y., Lilly, S. J., Renzini, A. & Carollo, C. M. Mass and environment as drivers of galaxy evolution. II. The quenching of satellite galaxies as the origin of environmental effects. *Astrophys. J.* **757,** 4 (2012).
32. Blanton, M. R. & Roweis, S. K-corrections and filter transformations in the ultraviolet, optical, and near-infrared. *Astron. J.* **133,** 734–754 (2007).
33. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344,** 1000–1028 (2003).
34. Brinchmann, J. *et al.* The physical properties of star-forming galaxies in the low-redshift Universe. *Mon. Not. R. Astron. Soc.* **351,** 1151–1179 (2004).
35. Gallazzi, A., Charlot, S., Brinchmann, J., White, S. D. M. & Tremonti, C. A. The ages and metallicities of galaxies in the local universe. *Mon. Not. R. Astron. Soc.* **362,** 41–58 (2005).
36. Lilly, S., Carollo, C. M., Pipino, A., Renzini, A. & Peng, Y. Gas regulation of galaxies: the evolution of the cosmic specific star formation rate, the metallicity-mass-star-formation rate relation, and the stellar content of halos. *Astrophys. J.* **772,** 119 (2013).
37. Peng, Y. & Maiolino, R. From haloes to galaxies—I. The dynamics of the gas regulator model and the implied cosmic sSFR history. *Mon. Not. R. Astron. Soc.* **443,** 3643–3664 (2014).
38. Bigiel, F. *et al.* The star formation law in nearby galaxies on sub-kpc scales. *Astrophys. J.* **136,** 2846–2871 (2008).
39. Savaglio, S. *et al.* The Gemini Deep Deep Survey. VII. The redshift evolution of the mass-metallicity relation. *Astrophys. J.* **635,** 260–279 (2005).
40. Maiolino, R. *et al.* AMAZE. I. The evolution of the mass-metallicity relation at $z > 3$. *Astron. Astrophys.* **488,** 463–479 (2008).
41. Petrosian, V. Surface brightness and evolution of galaxies. *Astrophys. J.* **209,** L1–L5 (1976).



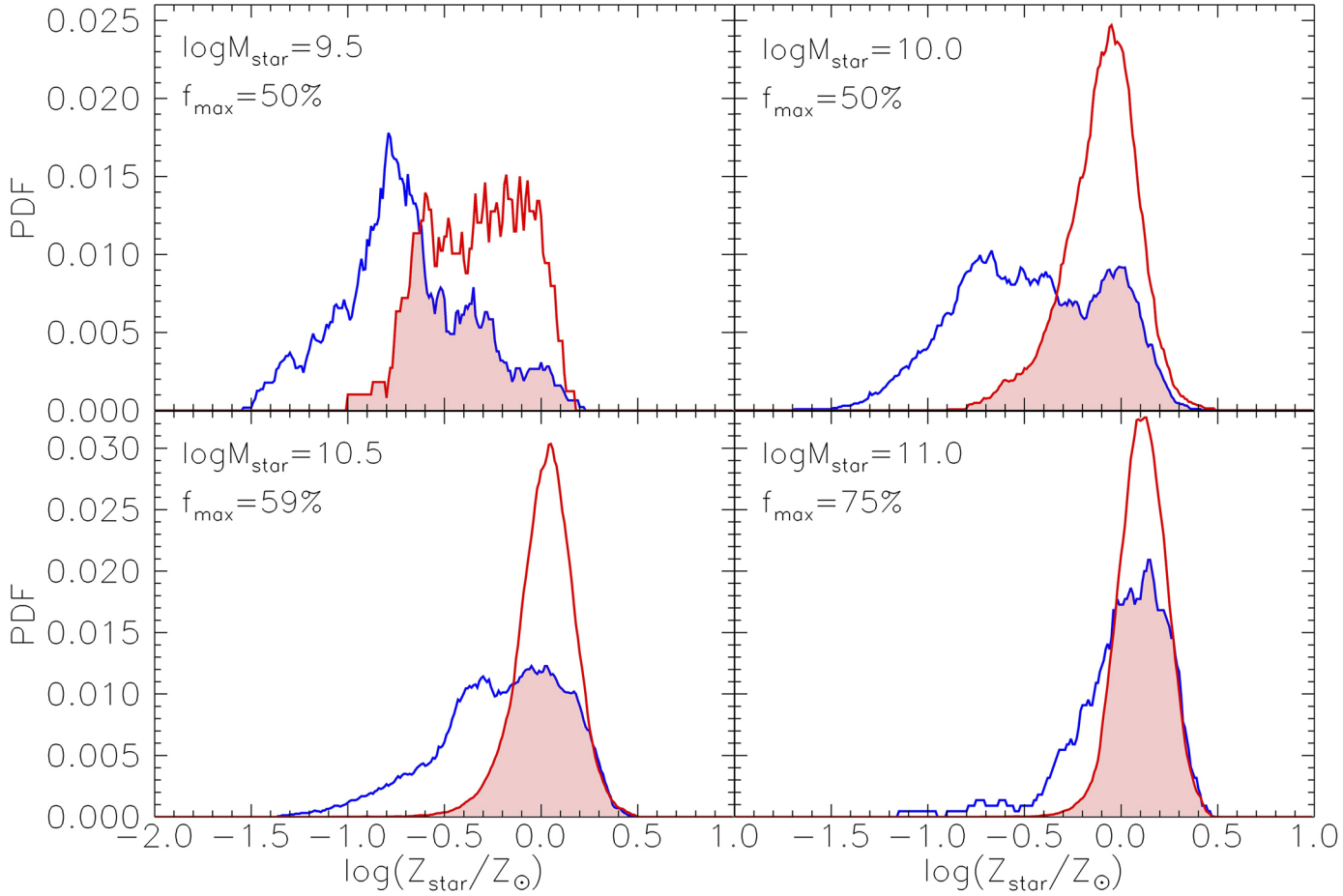
Extended Data Figure 1 | Gas fraction and star-formation efficiency for star-forming galaxies. **a**, The observed total gas fraction for local galaxies determined using a constant CO-to-H_2 conversion factor X_{CO} (black dashed line) and an H-band luminosity-dependent conversion factor (black solid line) in Boselli *et al.*[20]. The predicted total gas fraction (molecular and atomic) for star-forming galaxies as a function of stellar mass from the Peng and Maiolino[37] model at $z \approx 0.05$ (red solid line). **b**, Star-formation efficiency ε, defined as $\varepsilon = \mathrm{SFR}/M_{\mathrm{gas}}$, that is, the reverse of the gas depletion timescale, as a function of stellar mass.



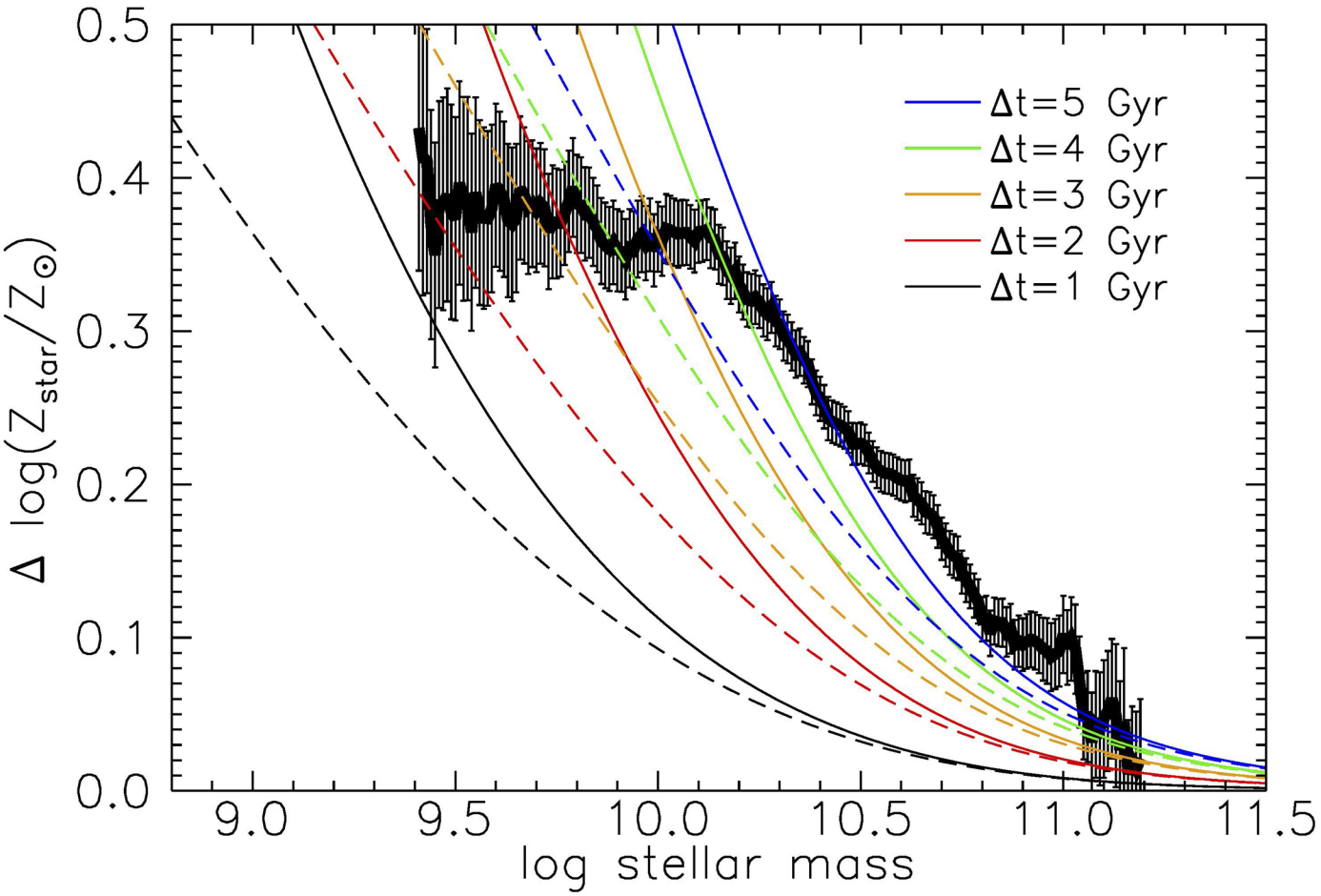
Extended Data Figure 2 | Stellar metallicity difference for different redshift bins. To investigate any aperture effects, the sample over the whole redshift range $0.02 < z < 0.085$ is further divided into two narrower redshift ranges of $0.02 < z < 0.05$ and $0.05 < z < 0.085$. It is clear that the derived stellar metallicity difference changes little as a function of redshift, that is, as a function of projected aperture. The error bars on each line indicate the 1σ uncertainty in the metallicity difference.



Extended Data Figure 3 | Stellar metallicity difference for central and satellite galaxies. The whole sample is further divided into central galaxies and satellites. The orange line shows the central galaxies in the field, which represents a clean sample of true central galaxies, as explained in the text. The stellar metallicity enhancement of satellites is slightly larger than that of central galaxies at $M_{\rm star} < 10^{10} M_{\odot}$ (suggesting that environment may play a part in the strangulation mechanism at these low masses), while no detectable difference between them is seen at higher stellar masses. The error bars indicate the 1σ uncertainty in the metallicity difference.



Extended Data Figure 4 | Probability density function of star-forming and passive galaxies. In each panel the blue line shows the probability density function (PDF) of the stellar metallicity of star-forming galaxies for a given stellar mass and the red line shows the corresponding PDF of passive galaxies. The overlapping region of the two PDFs is shaded (light red). The fraction of the shaded area over the total area given by each of PDF (f_{max}) gives the maximum fraction of galaxies for which rapid gas removal may be an allowed alternative quenching mechanism.



Extended Data Figure 5 | Effect of a constant star-formation efficiency on stellar metallicity evolution. As for Fig. 2b, but for the case of a constant star-formation efficiency of $\varepsilon = 0.5\,\mathrm{Gyr}^{-1}$ (that is, a constant gas depletion timescale of $\tau_{\mathrm{dep}} = 2$ Gyr) after strangulation. At $M_{\mathrm{star}} > \sim 10^{10} M_{\odot}$, the observed mass-dependent metallicity enhancement is still consistently $\Delta t \approx 4$ or 5 Gyr, while at lower stellar masses it requires a shorter Δt, as explained in the text. Error bars on the black line indicate the 1σ uncertainty in the metallicity difference.

LETTER

doi:10.1038/nature14398

Electron pairing without superconductivity

Guanglei Cheng[1,2], Michelle Tomczyk[1,2], Shicheng Lu[1,2], Joshua P. Veazey[1]†, Mengchen Huang[1,2], Patrick Irvin[1,2], Sangwoo Ryu[3], Hyungwoo Lee[3], Chang-Beom Eom[3], C. Stephen Hellberg[4] & Jeremy Levy[1,2]

Strontium titanate ($SrTiO_3$) is the first and best known superconducting semiconductor[1]. It exhibits an extremely low carrier density threshold for superconductivity[2], and possesses a phase diagram similar to that of high-temperature superconductors[3,4]—two factors that suggest an unconventional pairing mechanism. Despite sustained interest for 50 years, direct experimental insight into the nature of electron pairing in $SrTiO_3$ has remained elusive. Here we perform transport experiments with nanowire-based single-electron transistors at the interface between $SrTiO_3$ and a thin layer of lanthanum aluminate, $LaAlO_3$. Electrostatic gating reveals a series of two-electron conductance resonances—paired electron states—that bifurcate above a critical pairing field B_p of about 1–4 tesla, an order of magnitude larger than the superconducting critical magnetic field. For magnetic fields below B_p, these resonances are insensitive to the applied magnetic field; for fields in excess of B_p, the resonances exhibit a linear Zeeman-like energy splitting. Electron pairing is stable at temperatures as high as 900 millikelvin, well above the superconducting transition temperature (about 300 millikelvin). These experiments demonstrate the existence of a robust electronic phase in which electrons pair without forming a superconducting state. Key experimental signatures are captured by a model involving an attractive Hubbard interaction that describes real-space electron pairing as a precursor to superconductivity.

$SrTiO_3$ superconducts at temperatures below the superconducting transition temperature $T_c \approx 300$ mK and at electron densities[2] as low as 10^{17} cm^{-3}. Electrons can be introduced via doping (for example, Nb or La), oxygen vacancies[2] or electrolytic gating[5]. In 1969, it was predicted[6] that low-density superconductors (specifically, Zr-doped $SrTiO_3$) should exhibit unconventional real-space pairing in the absence of superconductivity. At higher temperatures, electrons were postulated to form tightly bound pairs; below the Bose–Einstein condensation (BEC) transition temperature, superconductivity was predicted to emerge.

The superconducting properties of $SrTiO_3$ have been previously investigated by electrical transport[1], tunnelling spectroscopy[7], and the Nernst effect[2]. New insights into the superconducting properties of $SrTiO_3$ come from heterointerfaces[8] that enable transport in reduced dimensions. The interface between TiO_2-terminated $SrTiO_3$ and a thin layer of $LaAlO_3$ supports a two-dimensional electron liquid[8] that exhibits electric-field-tunable superconductivity[3,9]. Recently, a pseudogap phase similar to that seen in high-T_c superconductors was observed at the $LaAlO_3/SrTiO_3$ interface using planar tunnelling spectroscopy[4].

The superconducting single-electron transistor (SET), consisting of an electrically gated superconducting quantum dot (QD) coupled to superconducting leads by tunnelling barriers, presents a particularly powerful tool for probing fundamental properties of superconductors[10]. Transport signatures of metallic superconducting islands include even–odd parity effects, Cooper pair tunnelling, and parity-affected superconductivity[11]. Generally, transport characteristics depend on the relative magnitudes of the charging energy E_c, superconducting gap energy Δ, and orbital level spacing δE in the QD.

Here we describe quantum transport measurements on $LaAlO_3/SrTiO_3$ SETs fabricated by conductive atomic force microscope lithography[12,13]. The devices are constructed from three basic elements: superconducting nanowires[14], nanoscale potential barriers created by conductive atomic force microscope erasure[15] and electrical side gates (see Methods). Figure 1a shows a schematic of a typical structure, consisting of a nanowire (between leads 1 and 5) of width $w \approx 5$ nm, three voltage probes (leads 2–4) and a side gate. Voltage leads are located a distance $L_w = 2.5$ μm apart, separating the main channel into two segments. The upper segment (between leads 2 and 3) is 'open', that is, without barriers, while the lower nanowire segment forms an $L_{QD} = 1$ μm QD bounded by two barriers. A side gate tunes the chemical potential of both the upper wire and the QD, and modulates the tunnel coupling between the QD and the external leads.

The low-temperature ($T = 50$ mK) differential conductance (dI/dV) versus side-gate voltage (V_{sg}) measurements for the nanowire QD and open wire show contrasting transport characteristics (device A, Fig. 1b, c). While the open wire (Fig. 1b) exhibits superconductivity[14] at all V_{sg} values shown, the QD (Fig. 1c) exhibits a sequence of diamond-shaped insulating regions for $V_{sg} < -10$ mV. The conductance increases by several orders of magnitude only when an available state in the QD is aligned within k_BT (k_B, Boltzmann's constant) of either the source or the drain chemical potential; this condition defines the diamond-shaped insulating regions in Fig. 1c. Within the diamonds, conductance through the QD is highly suppressed ($dI/dV < 10^{-2}\, e^2/h$, where e is the electronic charge, and h is Planck's constant). In the regime $V_{sg} > 0$ mV, where the barriers are highly transparent, supercurrent recovers and flows resonantly through the QD.

Generally, the 'addition energy' (the difference of chemical potentials μ_N and μ_{N+1}) required to change the charge state of a QD from N to $N+1$ electrons is the sum of both the classical charging energy E_c and the orbital energy δE of the device: $E_{add}(N) = E_c(N) + \delta E(N)$. For QD systems involving semiconductors, carbon nanotubes or superconductors[16], E_{add} is usually dominated by E_c, resulting in regularly spaced Coulomb diamonds. In device A, E_{add} decreases (non-monotonically) from 640 μeV (at $V_{sg} = -47$ mV) to 210 μeV (at $V_{sg} = -13$ mV). The level spacing is non-uniform, signifying that orbital contributions dominate the addition energy. Resonant supercurrent flowing through the QD is only observed when the addition energy E_{add} falls below the superconducting gap[4] $\Delta \approx 40$ μeV (for example, at $V_{sg} = -39$ mV and -19 mV), consistent with Anderson's criterion for nanoscale superconductivity ($\delta E < \Delta$)[17].

Figure 2a–e shows how the conductance diamonds evolve as a function of an applied out-of-plane magnetic field B. At $B = 0$ T, two zero-bias peaks (ZBPs) are visible, with some narrowing of the lineshape taking place at $B = 1$ T. The diamond pattern remains relatively unchanged at $B = 2$ T, though the size of the diamond is slightly reduced. At $B = 3$ T, new diamonds emerge and separate as the magnetic field is increased further to $B = 4$ T. A high-resolution scan

[1]Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. [2]Pittsburgh Quantum Institute, Pittsburgh, Pennsylvania 15260, USA. [3]Department of Materials Science and Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [4]Center for Computational Materials Science, Naval Research Laboratory, Washington DC 20375, USA.
†Present address: Department of Physics, Grand Valley State University, Allendale, Michigan 49401, USA.

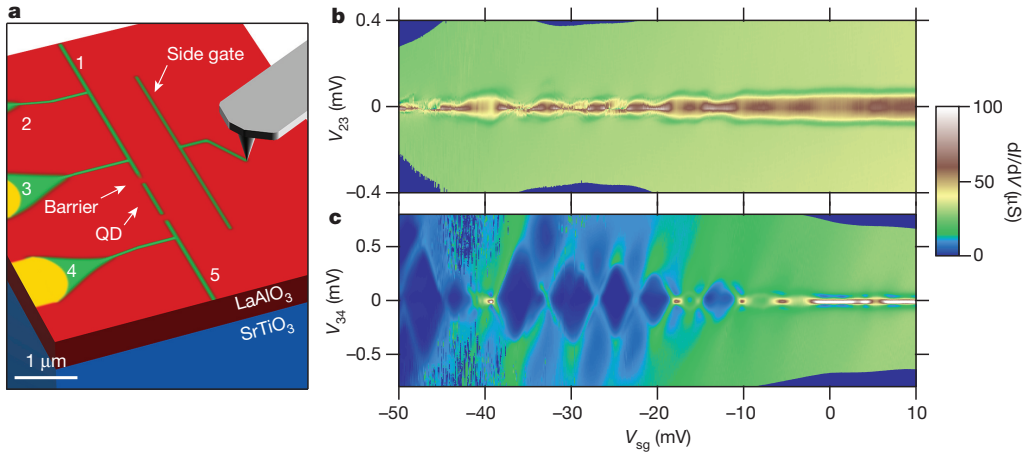


Figure 1 | Device schematic and transport characteristics. **a**, Device schematic. The nanowire width $w = 5$ nm, the nanowire QD length is 1 μm, and barriers are 0.75 μm away from the sense leads 3 and 4. The length of the open wire is 2.5 μm, equal to the nanowire QD length plus total distances from barrier to sense leads. **b**, Device A dI/dV characteristics (colour coded) as a function of four-terminal voltage V_{23} and side gate voltage V_{sg} in the open wire. **c**, dI/dV characteristics of the nanowire QD measured simultaneously with data shown in **b**.

of the conductance versus gate voltage at zero bias (Fig. 2f) enables the ZBP to be fitted and tracked versus magnetic field. A global shear of all of the ZBP splittings above B_p is observed and offset in Fig. 2f (see Extended Data Fig. 4). This shear, which appears in 60% of the total devices and is possibly attributable to orbital effects[11], does not influence the analysis of B_p. The ZBP at $V_{sg} = -27$ mV splits above a critical pairing field $B_p = 1.8 \pm 0.1$ T. The magnetic field at which this pairing transition occurs is one order of magnitude larger than the upper critical field for superconductivity (H_{c2}), $\mu_0 H_{c2} \approx 0.2$ T (here μ_0 is the vacuum permeability). The ZBPs at $V_{sg} = -19$ mV and $V_{sg} = -17$ mV have successively smaller values for B_p and show pronounced superconducting resonances below $|B| < 0.2$ T (indicated by the red arrow in Fig. 2f). For $|B| > B_p$, the energy difference between the split peaks increases linearly (Zeeman-like) with magnetic field

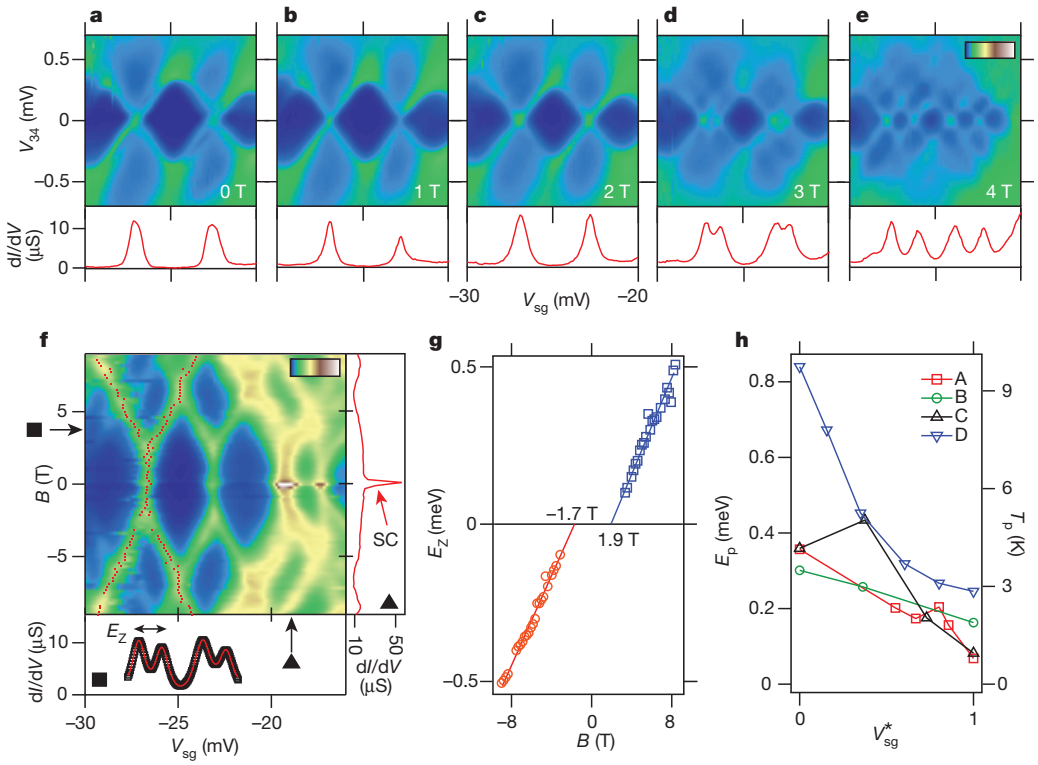


Figure 2 | Out-of-plane magnetic field dependence of device A transport characteristics at $T = 50$ mK. **a**–**e**, Top panels, dI/dV dependence on V_{34} and V_{sg} at $B = 0$ to 4 T. New diamonds emerge at $B = 3$ T in **d**. Colour scale (top right), 0–80 μS. Bottom panels, zero-bias line profiles in **a**–**e**. **f**, Top panel, magnetic field dependence of ZBPs. All the ZBPs split above some critical pairing fields, B_p. Colour scale, 0–40 μS. Bottom panel, line profile (black markers) of ZBP at $B = 3.8$ T in the top panel, indicated by the horizontal black arrow. Red line is the fit to extract peak locations (see Methods). Right panel, line profile at $V_{sg} = -19$ mV indicated by the vertical black arrow. The sharp peak at $B = 0$ is due to superconductivity (SC). **g**, Energy difference E_Z of two Zeeman splitting branches of the ZBP at $V_{sg} = -27$ mV. B_p and g factor can be extracted from the intercepts and slopes in the linear fits. **h**, E_p dependence on rescaled V^*_{sg} for all available ZBP splittings in four devices A, B, C and D, where $V^*_{sg} = \left(V_{sg} - V^{min}_{sg}\right) / \left(V^{max}_{sg} - V^{min}_{sg}\right)$, and V^{max}_{sg} and V^{min}_{sg} are maximum and minimum ZBP locations of each device. E_p roughly decreases with increasing V_{sg}.

$E_Z = g\mu_B(B - B_p)$. The Landé g-factor $g = 1.2 \pm 0.1$ (Fig. 2g) is calculated from the slope by taking into account the experimentally determined coupling factor $\alpha = 0.10 \pm 0.01\ \mathrm{eV\,V^{-1}}$ (Methods). At much larger magnetic fields the Zeeman-split ZBPs occasionally intersect and 'lock' together (re-entrant pairing) before separating again (for example, at $V_{sg} = -25$ mV). The energy associated with B_p, $E_p = g\mu_B B_p$, ranges between 100 μeV and 900 μeV for the four devices shown here (Fig. 2h) and decreases non-monotonically with increasing V_{sg} for each device.

Temperature-dependent transport measurements (Fig. 3) show B_p to be nearly independent of temperature up to the highest value measured ($T = 900$ mK). Of four devices, only device B, which shows the lowest B_p, exhibits a threefold suppression at $T = 900$ mK. Figure 3a shows a representative conductance map versus V_{sg} and B acquired at $T = 100$ mK. The conductance at $B = 6$ T shows two well-resolved split peaks at $T = 100$ mK. As the temperature is increased while the magnetic field is held constant at $B = 6$ T, the side-gate splitting between the two peaks ΔV_{sg} increases, shown in Fig. 3c. Assuming that the g-factor is constant over this temperature range, this result implies that $dB_p/dT < 0$, which is also consistent with the fitting result summarized in Fig. 3d.

The quantum transport behaviour of $LaAlO_3/SrTiO_3$ nanowire QDs contrasts sharply with conventional Coulomb blockade behaviour in other semiconductor nanostructures[16]. Experiments in Al-based superconducting SETs report pair tunnelling only in the superconducting state, and in the pair-tunnelling condition $E_c < 2\Delta$ when the pairing energy dominates over the Coulomb charging energy[11]. The high permittivity of $SrTiO_3$ leads to a significant reduction of the charging energy for $LaAlO_3/SrTiO_3$ nanowire SETs, enabling them to operate in the pair-tunnelling regime (see Methods for estimates and analysis). The observed ZBP splitting indicates that electron pairing is the preferred ground state that persists in magnetic fields far larger than $\mu_0 H_{c2}$, above which superconductivity is suppressed.

The existence of electron pairs outside the superconducting regime does not automatically imply that the electron pairing described here contributes to the superconductivity itself. It would, however, be a remarkable coincidence for the two phenomena to be superimposed without any interrelationship. In fact, electron pairing and superconductivity are demonstrably linked. The vertical linecut in Fig. 2f shows a sharp superconducting enhancement of the ZBP at $V_{sg} = -19$ mV. Like the other ZBPs, the paired electron state bifurcates at $B_p \approx 2$ T. This marked enhancement of conductance in the superconducting regime demonstrates that the electron pairs couple strongly to the superconducting leads.

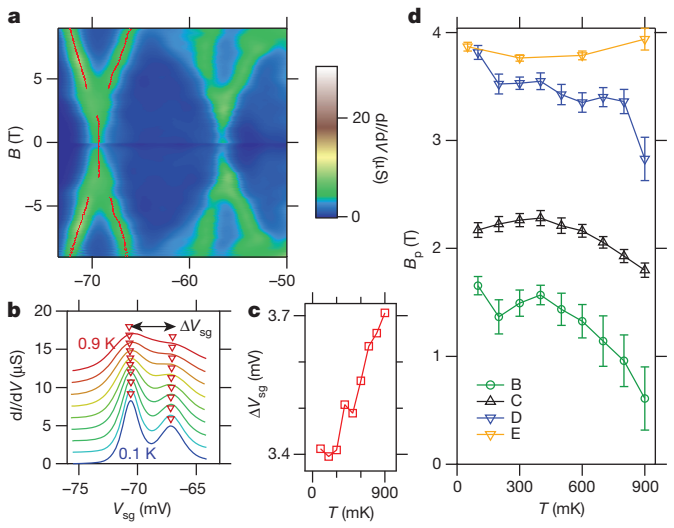


Figure 3 | Temperature dependence of B_p. **a**, Out-of-plane field dependence of ZBPs from device B at $T = 100$ mK. Trace of red dots reveals the actual peak locations extracted by fitting. **b**, Line profiles at $B = 6$ T of different temperatures from 100 mK (blue) to 900 mK (red) with 100 mK spacing. Red triangles mark actual peak positions and ΔV_{sg} is the difference between two splitting peak positions. Curves are offset for clarity. **c**, ΔV_{sg} in **b** as a function of temperature. A larger ΔV_{sg} at higher temperatures indicates lower B_p. **d**, Temperature dependence of B_p for the most isolated ZBP splittings in four devices B, C, D and E. B_p in device B, which is the lowest among the four devices, decreases non-monotonically with increasing temperature. Error bars, s.e.m. from the linear fitting errors of the positive and negative critical pairing fields.

Alternative explanations for the ZBP splittings have been considered. The Kondo ridge in Coulomb diamonds can split above a critical magnetic field[18]. However these splittings are generally observed at non-zero biases; furthermore, other Kondo parity signatures[19] are not observed here. Charge traps that exist in parallel with tunnelling barriers can release additional electrons to the transport[20], resulting in occasional resonance-doubling features (Methods). The main features reported here are consistently reproduced in more than 50 devices (see Extended Data Fig. 6 for more examples), and do not fit the statistical profile of charge traps. In ultrasmall superconducting grains where $\delta E \gg \Delta$, quantum fluctuations may promote the even–odd parity energy, leading to a possibly similar ZBP splitting[21]. Such an effect, however, is only expected for $T < T_c$.

Electron pairing without superconductivity can be described by a phenomenological Fermi–Hubbard model (equation (1)) with an attractive on-site potential[22]. The QD is represented by a one-dimensional chain of local pairing sites that can be occupied with zero, one, or two electrons. The Hamiltonian is written as

$$H = -t\sum_{i,\sigma}\left(c^{+}_{i+1,\sigma}c_{i,\sigma} + c^{+}_{i,\sigma}c_{i+1,\sigma}\right) + U\sum_{i} n_{i\uparrow}n_{i\downarrow} + g\mu_B B\sum_{i} S^{z}_{i} - \mu\sum_{i,\sigma} n_{i,\sigma} \quad (1)$$

where $c^{+}_{i,\sigma}$ and $c_{i,\sigma}$ are creation and annihilation operators for electrons on site i with spin $\sigma = \uparrow,\downarrow$; t quantifies the effective hopping between adjacent pairing sites; $n_{i,\sigma} = c^{+}_{i,\sigma}c_{i,\sigma}$ is the number operator; $S^{z}_{i} = (1/2)\left(n_{i,\uparrow} - n_{i,\downarrow}\right)$ is the spin operator; $U < 0$ represents the on-site attractive interaction strength; B is the applied magnetic field; and μ is the chemical potential. For two electrons, equation (1) can be solved exactly (Methods), yielding analytic expressions for the pairing energy $\Delta_p(B) = \sqrt{16t^2 + U^2} - 4t - g\mu_B B$ and critical pairing field $B_p = \left(-4t + \sqrt{16t^2 + U^2}\right)/g\mu_B$.

The zero-temperature stability diagram of the 16-site model (Fig. 4b) qualitatively captures many of the experimentally observed features: the existence of a critical pairing field B_p, a Zeeman-like splitting for $|B| > B_p$, a decrease of B_p with increasing μ, and re-entrant pairing at higher magnetic fields (Fig. 2f). The Hamiltonian (equation (1)) has no disorder, resulting in the even level spacing seen in Fig. 4b. Adding some disorder to the Hamiltonian, either in the energy levels of each pairing site or in the kinetic hopping between pairing sites, makes the level spacings less regular, more closely resembling the spacings seen in the experimental ZBPs (Fig. 4a). Additionally, the superconducting regime of the attractive Hubbard model is not explored here, but is covered extensively in the literature[23]. Being phenomenological, equation (1) does not specify a physical mechanism for the attractive on-site interaction. One form of the attractive interaction can be 'negative-U' centres, for example, oxygen vacancies or vacancy complexes, similar to those which have been linked to enhanced superconducting transition temperatures in Tl-doped PbTe (refs 24, 25). Bipolaronic mechanisms have also been advocated in the literature, especially by Alexandrov[26] (Methods). Such specific mechanisms for strong pairing are not directly implied by the measurements reported here, although some predictions may be testable with suitable refinements of this experimental approach.

Note that spin–orbit coupling is neglected in this analysis, even though such effects are known to be important in two-dimensional

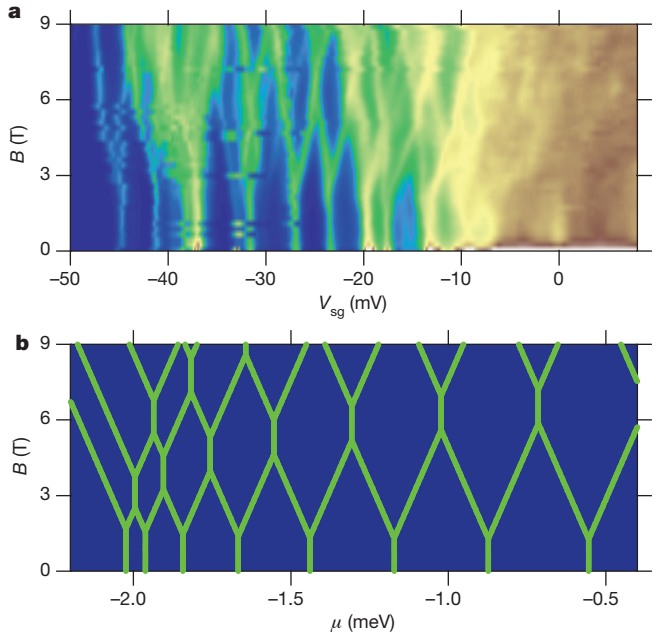


Figure 4 | Comparison between experiment and attractive Hubbard model. **a**, Dependence of ZBPs on V_{sg} and B of device A on a larger V_{sg} scale (same colour scheme as Fig. 2f with a scale of 0–50 μS). **b**, Simulation result from the attractive Hubbard model of 16 sites with $t = 1$ meV and $U = -0.8$ meV.

transport experiments[27,28]. Spin–orbit coupling makes electron pairs less sensitive to magnetic fields and leads to the violation of the Pauli limit in $SrTiO_3$ (Pauli limiting field $\mu_0 H_c^P = 1.84 T_c \approx 0.5$ T)[27,29]. However, it is not clear how such coupling will increase the pairing energy above T_c. At the $LaAlO_3/SrTiO_3$ interface, spin–orbit coupling is known[27,28] to be strongly dependent on the carrier density n_s. Direct measurements of carrier density are not feasible in the geometry employed here, although the density is believed to increase monotonically with gate voltage.

The existence of pre-formed electron pairs in this $SrTiO_3$-based system, forming a superconducting condensate at lower temperatures and lower magnetic fields, follows the paradigm of BEC superconductivity. In the BEC regime, pairing is local and precedes the formation of a superconducting state. The only well-established physical embodiments of fermionic BEC-like superfluidity have been in ultracold atomic gases, where the BEC–BCS (Bardeen–Cooper–Schrieffer) crossover can be tuned via a Feshbach resonance[30]. Although it is not clear whether the strength of the electron pairing can be tuned (for example, via strain), a crossover to BCS-like superconductivity at higher electron density is expected. The ability to confine electrons at nanoscale dimensions, combined with an inherent affinity for strong pairing, suggests that our system constitutes an ideal 'laboratory' in which to explore strongly correlated electronic phases in a solid-state host.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 October 2014; accepted 4 March 2015.

1. Schooley, J. F., Hosler, W. R. & Cohen, M. L. Superconductivity in semiconducting $SrTiO_3$. *Phys. Rev. Lett.* **12,** 474–475 (1964).
2. Lin, X., Zhu, Z., Fauqué, B. & Behnia, K. Fermi surface of the most dilute superconductor. *Phys. Rev. X* **3,** 021002 (2013).
3. Caviglia, A. D. *et al.* Electric field control of the $LaAlO_3/SrTiO_3$ interface ground state. *Nature* **456,** 624–627 (2008).
4. Richter, C. *et al.* Interface superconductor with gap behaviour like a high-temperature superconductor. *Nature* **502,** 528–531 (2013).
5. Ueno, K. *et al.* Electric-field-induced superconductivity in an insulator. *Nature Mater.* **7,** 855–858 (2008).
6. Eagles, D. M. Possible pairing without superconductivity at low carrier concentrations in bulk and thin-film superconducting semiconductors. *Phys. Rev.* **186,** 456–463 (1969).
7. Binnig, G., Baratoff, A., Hoenig, H. E. & Bednorz, J. G. Two-band superconductivity in Nb-doped $SrTiO_3$. *Phys. Rev. Lett.* **45,** 1352–1355 (1980).
8. Ohtomo, A. & Hwang, H. Y. A high-mobility electron gas at the $LaAlO_3/SrTiO_3$ heterointerface. *Nature* **427,** 423–426 (2004).
9. Reyren, N. *et al.* Superconducting interfaces between insulating oxides. *Science* **317,** 1196–1199 (2007).
10. von Delft, J., Zaikin, A. D., Golubev, D. S. & Tichy, W. Parity-affected superconductivity in ultrasmall metallic grains. *Phys. Rev. Lett.* **77,** 3189–3192 (1996).
11. Tinkham, M., Ralph, D. C., Black, C. T. & Hergenrother, J. M. Discrete energy levels and superconductivity in nanometer-scale Al particles. *Czech. J. Phys.* **46,** 3139–3145 (1996).
12. Cen, C. *et al.* Nanoscale control of an interfacial metal-insulator transition at room temperature. *Nature Mater.* **7,** 298–302 (2008).
13. Cen, C., Thiel, S., Mannhart, J. & Levy, J. Oxide nanoelectronics on demand. *Science* **323,** 1026–1030 (2009).
14. Veazey, J. P. *et al.* Oxide-based platform for reconfigurable superconducting nanoelectronics. *Nanotechnology* **24,** 375201 (2013).
15. Cheng, G. L. *et al.* Sketched oxide single-electron transistor. *Nature Nanotechnol.* **6,** 343–347 (2011).
16. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
17. Anderson, P. W. Theory of dirty superconductors. *J. Phys. Chem. Solids* **11,** 26–30 (1959).
18. Costi, T. A. Kondo effect in a magnetic field and the magnetoresistivity of Kondo alloys. *Phys. Rev. Lett.* **85,** 1504–1507 (2000).
19. Goldhaber-Gordon, D. *et al.* Kondo effect in a single-electron transistor. *Nature* **391,** 156–159 (1998).
20. Hofheinz, M. *et al.* Individual charge traps in silicon nanowires — measurements of location, spin and occupation number by Coulomb blockade spectroscopy. *Eur. Phys. J. B* **54,** 299–307 (2006).
21. Matveev, K. A. & Larkin, A. I. Parity effect in ground state energies of ultrasmall superconducting grains. *Phys. Rev. Lett.* **78,** 3749–3752 (1997).
22. Anderson, P. W. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235,** 1196–1198 (1987).
23. Micnas, R., Ranninger, J. & Robaszkiewicz, S. Superconductivity in narrow-band systems with local nonretarded attractive interactions. *Rev. Mod. Phys.* **62,** 113–171 (1990).
24. Matsushita, Y., Bluhm, H., Geballe, T. H. & Fisher, I. R. Evidence for charge Kondo effect in superconducting Tl-doped PbTe. *Phys. Rev. Lett.* **94,** 157002 (2005).
25. Dzero, M. & Schmalian, J. Superconductivity in charge Kondo systems. *Phys. Rev. Lett.* **94,** 157003 (2005).
26. Alexandrov, A. S. in *Polarons in Advanced Materials* (ed. Alexandrov, A. S.) Ch. 7 257–310 (Springer Series in Materials Science, Vol. 103, Springer, 2007).
27. Ben Shalom, M., Sachs, M., Rakhmilevitch, D., Palevski, A. & Dagan, Y. Tuning spin-orbit coupling and superconductivity at the $SrTiO_3/LaAlO_3$ interface: a magnetotransport study. *Phys. Rev. Lett.* **104,** 126802 (2010).
28. Caviglia, A. D. *et al.* Tunable Rashba spin-orbit interaction at oxide interfaces. *Phys. Rev. Lett.* **104,** 126803 (2010).
29. Kim, M., Kozuka, Y., Bell, C., Hikita, Y. & Hwang, H. Y. Intrinsic spin-orbit coupling in superconducting delta-doped $SrTiO_3$ heterostructures. *Phys. Rev. B* **86,** 085121 (2012).
30. Zwierlein, M. W., Abo-Shaeer, J. R., Schirotzek, A., Schunck, C. H. & Ketterle, W. Vortices and superfluidity in a strongly interacting Fermi gas. *Nature* **435,** 1047–1051 (2005).

Acknowledgements We thank A. Akhmerov, A. Annadi, S. Frolov, R. Lutchyn, C. Nayak and D. Pekker for discussions. This work was supported by ARO MURI W911NF-08-1-0317 (J.L.), AFOSR MURI FA9550-10-1-0524 (C.-B.E., J.L.) and FA9550-12-1-0342 (C.-B.E.), grants from the National Science Foundation DMR-1104191 (J.L.), DMR-1124131 (C.-B.E., J.L.) and DMR-1234096 (C.-B.E.), and the Office of Naval Research through the Naval Research Laboratory's Basic Research Program (C.S.H.).

Author Contributions G.C. and M.T. did most of the design and fabrication of the devices, performed the experiments and wrote the manuscript. S.L. and J.P.V. contributed to measurements. M.H. patterned the interface electrodes. P.I. contributed to manuscript writing and measurement set-up. S.R. and H.L. grew the samples. C.-B.E. supervised sample growth and reviewed the manuscript. C.S.H. performed theoretical calculations and co-wrote the manuscript. J.L. supervised all the related experiment procedures and co-wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L. (jlevy@pitt.edu).

METHODS

Sample growth. $LaAlO_3/SrTiO_3$ samples are grown by pulsed laser deposition (PLD). The $SrTiO_3$ substrate is TiO_2 terminated by etching in buffered HF for 60 s either once or twice to improve substrate quality. Then the $SrTiO_3$ substrate is annealed at 1,000 °C for 2–12 h to achieve an atomically smooth surface. A thin (3.4 unit cell) $LaAlO_3$ film is subsequently grown on top of $SrTiO_3$ by PLD at a temperature of 550 °C and 1×10^{-3} mbar oxygen pressure, and gradually cooled down to room temperature. Electrical contact to the $LaAlO_3/SrTiO_3$ interface is made by Ar^+ etching (25 nm) followed by sputter deposition of Ti/Au (5 nm/20 nm). Additional details are given in ref. 15.

Device fabrication. The conductive atomic force microscope fabrication technique[13] provides great versatility in the creation of nanoscale devices[31,32]. The key device element is a nanoscale tunnel barrier that is integral to many devices, including nanowire junctions[13], sketched field-effect transistors ('SketchFETs')[13], photodiodes[33], THz emitters and detectors[34], and sketched single-electron transistors ('SketchSETs')[15]. In this work, conductive atomic force microscope lithography is used to create 58 QD devices with varying dimensions (for example, $250\,\text{nm} < L_{QD} < 1\,\mu\text{m}$) and barrier heights ΔR (see definition below) and on multiple $SrTiO_3$ substrates. Every device exhibits features that are qualitatively similar to device A. In the following, a single barrier device, which has a similar design to device A but contains only one barrier instead of two, is shown in order to demonstrate the device fabrication technique (Extended Data Fig. 1a). During the barrier fabrication, the four-terminal resistance difference $\Delta R = R_{QD} - R_O$ is monitored in real time and serves as a figure of merit for low-temperature barrier performance, where R_{QD} and R_O are the resistances of the QD wire and the open wire. R_{QD} and R_o are obtained simultaneously by two four-terminal measurements using two hardware-simulated lock-in amplifiers, such that wire decay is eliminated from the measurement and ΔR is very precise. Prior to barrier cutting, R_{QD} and R_O are nominally the same with $\Delta R < 5\,\text{k}\Omega$ (within 1% difference). A sharp AFM tip (nominal radius of curvature ~8 nm) moves across the wire at 200 nm s^{-1} speed with small negative voltages (−0.1 V to −0.5 V) multiple times, which causes ΔR to increase discretely, as shown in Extended Data Fig. 1b. Low-temperature transport study ($T = 75$ mK) shows the wire conductance can be pinched off by V_{sg} for a single barrier device with $\Delta R \approx 120\,\text{k}\Omega$. The conductance oscillations ($30\,\text{mV} < V_{sg} < 50\,\text{mV}$) indicate quantum mechanical tunnelling through the barrier (Extended Data Fig. 1c).

Barrier height. The barrier resistance ΔR is a good indicator of the tunnel barrier width and strongly influences the low-temperature transport properties. Extended Data Fig. 2 shows transport characteristics of three 500 nm QD devices (devices F, G and H). The device designs are similar to that of device A, but with a shorter distance of 500 nm between the two barriers. The only difference among devices F, G and H is the single barrier resistance $\Delta R/2$, with $\Delta R/2 = 20$, 110 and 305 kΩ respectively. The resulting difference in transport is clear. Device F shows the most conductance diamonds for the smallest barrier resistance (superconductivity is suppressed in this device after applying a V_{bg} of −5.6 V), while device H, with the largest barrier resistance, is virtually featureless. Conductance diamonds and superconductivity-related phenomena in dI/dV of device G have V_{sg}-dependence that is very similar to that of device A. Since the single barrier tunnelling rate decays exponentially with the barrier width, the differences in transport of devices F, G and H can be understood as a result of suppressed quantum tunnelling rates under the assumption that the two barriers are slightly asymmetric.

Additional data for device A. The barriers can fully suppress the conductance of the wire in the low-V_{sg} regime, while allowing resonant tunnelling current at selected low V_{sg} values, and fully superconducting transport at high V_{sg} values. At low V_{sg} values, the line cut in Extended Data Fig. 3a shows ZBPs separated by regions of fully suppressed conductance. At high V_{sg} values, conductance oscillations due to quantum tunnelling are reduced as the barriers become increasingly transparent and the device enters a fully superconducting state.

Below B_p, the ZBPs are generally insensitive to magnetic fields. Above B_p, the centres of the ZBP splittings move nonlinearly with magnetic fields, as shown in the red trace $\delta V_{sg}(B)$ in Extended Data Fig. 4a. This movement, possibly arising from the orbital effect, is only observed in a fraction of devices. A similar effect has been reported in ref. 35. This is corrected (Extended Data Fig. 4b) by offsetting the magnetically induced global shift to keep the centres of ZBP splittings relatively constant, that is, $V_{sg}(B) = V_{sg0}(B) - \delta V_{sg}(B)$. This shift procedure does not influence the analysis of B_p, nor does it change the relative spacing between splittings. A waterfall plot (Extended Data Fig. 5) of Fig. 2f is included to provide clear line cuts of the ZBPs at all fields.

Analysis of B_p. Above B_p, single electron tunnelling occurs and standard Coulomb blockade physics analysis applies. Closely spaced Coulomb blockade ZBPs can be fitted to a multi-peak hyperbolic cosine expression[36]

$$dI/dV = G_0 + \sum_i A_i \cosh^{-2}\left[B_i\left(V_{sg} - V_{sg}^i\right)\right] \quad (2)$$

where G_0, A_i and B_i are constants. Equation (2) is used to examine two peaks that have split from a single ZBP at low magnetic fields in order to extract the ith peak location as a function of field ($V_{sg}^i(B)$ indicated in the red trace in the main panel of Fig. 2f in the main text). This fit was performed for multiple pairs of split peaks in many devices. The energy difference of the split peaks, $E_Z(B) = \alpha(V_{sg}^2(B) - V_{sg}^1(B))$, where α is the coupling factor as described in the next section, can be fitted to a straight line as in Fig. 2g. The slope of this fit gives the Landé g-factor according to $E_Z = g\mu_B(B - B_p)$, where the Zeeman energy difference is offset by the pairing energy $E_p = g\mu_B B_p$. As mentioned in the last section, 'straightening' the magnetically induced global shift to keep the centres of ZBP splittings relatively constant does not disturb this analysis. Both peak positions $V_{sg}^1(B)$ and $V_{sg}^2(B)$ have the same offset at each field, so E_Z(B), from which g and B_p are calculated, remains unaffected.

Device transport parameters. As mentioned in the main text, bulk $SrTiO_3$ is an incipient ferroelectric at low temperatures with a divergently large and gate-tunable dielectric constant[37] $\varepsilon_r \approx 20{,}000$. Consequently, estimation of the gate–dot capacitance yields $C_{sg} \approx 100\,\text{fF}$, using a parallel wire model $C_{sg} = \pi\varepsilon_r\varepsilon_0 L/\cosh^{-1}(d/2r)$ with vacuum permittivity ε_0, QD length $L = 1\,\mu\text{m}$, QD–gate spacing $d = 1\,\mu\text{m}$, and nanowire radius $r_{QD} = 5\,\text{nm}$ = side gate radius $r_{sg} = 5\,\text{nm}$. The corresponding charging energy is vanishingly small in the ideal case (that is, zero V_{sg}) $E_c = e^2/C_\Sigma < e^2/C_{sg} = 2\,\mu\text{eV}$, where C_Σ is the total capacitance of the QD. The actual E_c could be larger since ε_r is expected to be reduced by electric field and strain effects at the interface.

For a nanowire with length $L = 1\,\mu\text{m}$ and width $w = 5\,\text{nm}$, the number of carriers can be estimated as $N = n_s L w = 500$ by using a typical two-dimensional $LaAlO_3/SrTiO_3$ carrier density $n_s \approx 10^{13}\,\text{cm}^{-2}$. The mean level spacing between spin-degenerate levels can be estimated by using a 'particle in a one-dimensional box' model and effective mass[38] $m^* = 0.7\,m_e$. This gives $\delta E = \frac{\partial E}{\partial N} = \frac{\pi^2 N\hbar^2}{m^* L^2} \approx 500\,\mu\text{eV}$, which is consistent with the values of E_{add}. As one can see, the addition energy is dominated by orbital level spacing δE.

The ability of the side gate to tune the chemical potential of the device is characterized by $\alpha = C_{sg}/C_\Sigma$. The coupling factor can be calculated using the slopes β and γ which define the diamonds in the dI/dV map, $1/\alpha = 1/\beta + 1/\gamma$. Coupling factors vary from $\alpha \approx 0.03$ to 0.13, with a typical $\alpha \approx 0.10$. For all devices, the coupling factor is observed to decrease at high V_{sg} values; this variation is reflected in Extended Data Table 1.

Constant interaction model. The constant interaction model is widely used to analyse QD transport characteristics through two independent variables: Coulomb interactions and single-particle energy levels[16]. Here, superconductivity is combined with the constant interaction model and the analysis from ref. 39 is extended by including non-zero orbital level spacing. In a QD with N electrons, the excess charge has two parts: the integer part $n = N - N_0$ and a continuous part $C_{sg}V_{sg}/e$ representing electrostatic charge induced by the gate, where N_0 is the charge at zero gate voltage. The system ground state energy $E(N)$ can be written as

$$E(N) = \sum_{i=1}^{N} E_i + E_c\left(n - V_{sg}\alpha e/E_c\right)^2\Big/2 + p\delta_p \quad (3)$$

where E_i are single-particle energy levels, p is a parity factor with $p = 0$ (1) for even (odd) N and δ_p is parity energy. The first term in equation (3) is the electrochemical contribution determined by quantum confinement, the second term is electrostatic part induced by V_{sg} and the third term is the extra energy $\delta_p = E(N_{odd}) - (E(N_{odd}+1) + E(N_{odd}-1))/2$ which the odd electron has to pay to enter the QD. The addition energy E_{add}, which is the difference (of chemical potential μ) of a difference (of total energy E), is directly measured in the tunnelling spectroscopy measurement. Namely, the chemical potential is

$$\mu(N) = E(N) - E(N-1) = E_N + E_c(n - 1/2) - e\alpha V_{sg} + \beta\delta_p \quad (4)$$

where $\beta = -1$ (1) for even (odd) number of electrons (N). E_{add} can subsequently be written as

$$E_{add} = \mu(N+1) - \mu(N) = E_c + \delta E(N) + \gamma\delta_p \quad (5)$$

where $\gamma = 2$ (−2) for even (odd) number of electrons. Interestingly, as described in the main text, $E_{add} = E_c - 2\delta_p$ can be negative in the odd case ($\delta E(N_{odd}) = 0$) if $E_c < 2\delta_p$, suggesting this unpaired electron is not stable and wants to pair with a partner. In a BCS superconductor, the parity energy is approximately the

gap energy ($\delta_p \approx \Delta$) in the limit of small level spacing $\delta E(N)$ compared to the superconducting gap Δ ($\delta E < \Delta$). In the opposite extreme limit $\delta \gg \Delta$, δ_p can be enhanced such that $\delta_p = \delta E/2\ln(\delta E/\Delta)$ due to quantum fluctuations[21]. Either way it is reasonable to assume $E_c < 2\delta_p$ based on the estimate of E_c, suggesting pair tunnelling is the preferred transport mechanism in the devices explored here. Note in the case of pairing without superconductivity, the parity energy δ_p should be replaced by the pair binding energy Δ_b in equations (3)–(5). When the temperature and magnetic field are increased, δ_p and Δ_b are suppressed to zero at the same B_p where the Zeeman splitting of the peaks occurs.

Attractive Hubbard model. In order to obtain more insights into the pairing picture, the analysis is further extended by using the attractive Hubbard model, which qualitatively captures many features seen in the devices. The Hamiltonian is written as in equation (1) in the main text. The only assumption is the attractive pairing potential $U < 0$. The kinetic hopping $t > 0$ describes the motion of the electron. For sufficiently attractive $U < 0$, electrons will bind into pairs. This is the regime of the parity effect: in zero external field, the ground state as a function of gate voltage (represented by the chemical potential in equation (1)) will always contain an even number of electrons. The external magnetic field B favours polarized states, breaking the pairs. For magnetic fields greater than a critical field $B > B_p$, ground states with odd electron numbers can be stabilized. The interaction between pairs causes the critical field to decrease monotonically with increasing filling (increasing chemical potential or gate voltage).

As a simple example, the Hubbard model is solved on an infinite chain with zero, one and two electrons. The Bethe ansatz may be used to solve the model for arbitrary filling, but it is much more complicated than the approach presented here[40]. For zero and one electrons, the energies are simply given by:

$$E_0 = 0$$

$$E_1 = -2t - \frac{1}{2}g\mu_B B - \mu$$

Only the low-field two-electron ground state, which is a spin singlet[41,42], is considered for this analysis. The triplet will be the ground state at higher fields for two electrons. The ground state has zero momentum, so the wavefunction depends only on the separation between the electrons and must behave exponentially. Thus the non-normalized wavefunction is

$$\phi(i,j) = e^{-\zeta|i-j|}$$

for electrons on sites i and j. For $i \neq j$, Schrodinger's equation gives

$$E_2 = -2t\left(e^{-\zeta} + e^{\zeta}\right) - 2\mu \quad (6)$$

while for $i = j$ it gives

$$E_2 = U - 4te^{-\zeta} - 2\mu \quad (7)$$

Combining equations (6) and (7) yields

$$\zeta = \log\left(\frac{-U + \sqrt{16t^2 + U^2}}{4t}\right)$$

$$E_2 = -\sqrt{16t^2 + U^2} - 2\mu$$

Thus the binding energy for an electron pair is

$$\Delta_b = 2E_1 - E_2 = \sqrt{16t^2 + U^2} - 4t - g\mu_B B$$

and the 'size' of the pair is simply $1/\zeta$. The boundary between the phases with 0 and 1 electrons is given by

$$B_{0,1} = -\frac{4t + 2\mu}{g\mu_B}$$

The boundary between the phases with 1 and 2 electrons is given by

$$B_{1,2} = \frac{-4t + 2\sqrt{16t^2 + U^2} + 2\mu}{g\mu_B}$$

and the boundary between the phases with 0 and 2 electrons is independent of B:

$$\mu_{0,2} = -\frac{1}{2}\sqrt{16t^2 + U^2}$$

The plot of these three boundaries will have the shape of the letter 'Y'. The three boundaries meet at a critical point given by

$$\mu_p = -\frac{1}{2}\sqrt{16t^2 + U^2}$$

$$B_p = \frac{-4t + \sqrt{16t^2 + U^2}}{g\mu_B}$$

To expand the discussion, the lowest eigenvalues of the Hubbard Hamiltonian (equation (1)) are solved on a 16-site chain using the iterative Lanczos algorithm, which is particularly effcient for sparse matrices[43–46]. The full Hilbert space has 4^{16} states, which are split into smaller subspaces using the total electron number, $N_e = \sum_{i,\sigma} n_{i,\sigma}$, the total z-component of spin, $S^z = \sum_i S_i^z$, and the mirror symmetry of the system. The total spin is an additional symmetry of the Hamiltonian which is not exploited. The largest subspace contains 82,820,900 states.

Extended Data Fig. 7a shows the energy of the ground state for $N_e \leq 16$ in zero applied magnetic field. The lowest-energy state has total $S^z = 0$ for N_e even and $S^z = 1/2$ for N_e odd. The ground state always has an even number of electrons, which can be seen by shifting all the energies by a suitable function of μ, chosen here as quadratic. This does not change their relative order but can make the energy differences easier to visualize. In Extended Data Fig. 7b, the parity effect is apparent: the ground state always contains an even number of electrons.

Increasing the magnetic field reduces the energy of the higher spin states relative to the $S = 0$ ground states at $B = 0$. The ground state is polarized in the $-z$ direction, so the total spin S is identical to the z-component of spin S^z. The phase diagram as a function of magnetic field and chemical potential μ is shown in Extended Data Fig. 8. At low magnetic fields, the system consists of electron pairs: N_e is even, and the total spin $S = 0$. At slightly higher fields, it becomes favourable to have a single unpaired electron, resulting in odd N_e and $S = 1/2$. Increasing the field further results in two unpaired electrons. Now N_e is even again, but with total spin $S = 1$. The pattern continues with increasing field—the number of unpaired electrons increases monotonically.

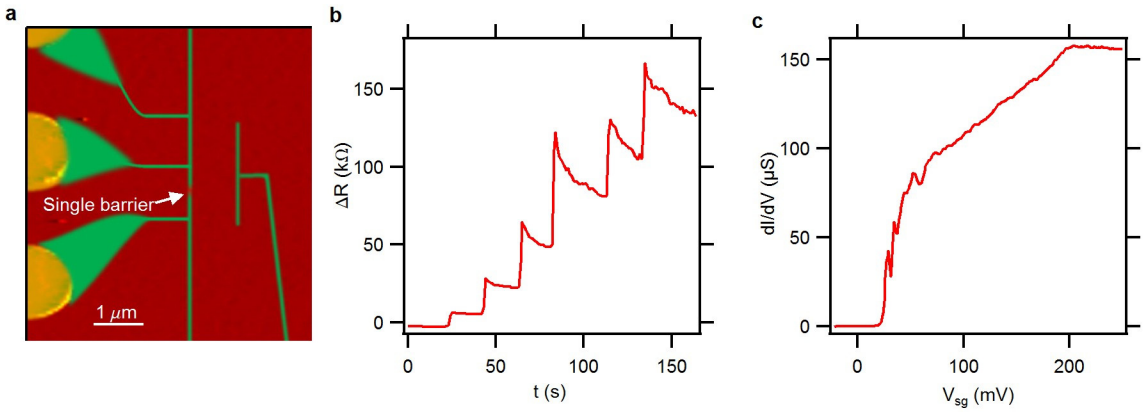
Pairing mechanisms. The attractive Hubbard model does not specify a physical origin of the pairing mechanism. Here we discuss two possible forms of pairing sites: negative-U centres and bipolarons. The negative-U centre, which hosts a bounded electron pair, was first proposed in ref. 47 to account for the diamagnetism in amorphous semiconductors. Its existence has been reported in various materials, for example, hydrogenic[48] or oxygen impurities[49] in GaAs. Meanwhile, negative-U centres have been proposed as a pairing mechanism in some unconventional superconductors such as Tl-doped PbTe (refs 24, 25). In $SrTiO_3$, negative-U centres can possibly originate from oxygen vacancies (or vacancy clusters) since the lowest threshold carrier density is only observed in vacancy doped samples but not in samples with other n-type dopants (for example, Nb)[50]. Another possible mechanism for local pairing is bipolaron formation. Bipolarons are bound states of two polarons[23], which are self-localized electronic states formed from lattice distortions—for example, via the Jahn-Teller effect[51]. When two polarons meet, they can share the same lattice distortion, lowering the total energy per electron and thus forming a bound state under certain conditions. The existence of polarons in $SrTiO_3$ has been extensively reported (see, for example, refs 52, 53). While there is no definitive experimental evidence of bipolaron formation in $SrTiO_3$, there are reports of bipolarons in other titanites[54].

Single-electron charge traps. A single-electron charge trap can be modelled with a series of capacitances that reflect the coupling between the trap and the QD, source and drain, as described in ref. 20. When the trap is in series with the QD, a large source–drain bias is needed to allow electrons to pass through the typically misaligned energy levels of the trap and QD. Namely, the conductance diamonds will have a large gap close to the zero-bias region (in contrast to our observations). When the trap is in parallel with the QD, the contribution to the conductance will be negligible since the coupling between the trap and either source or drain will be very weak due to the small trap size (compared to the 1 μm nanowire QD length). A more realistic scenario is a combination of both the series and parallel coupling. Namely, the trap is in parallel with one of the tunnel barriers and can occasionally release an electron to QD, which is commonly referred as the background or offset charge[55,56]. The transport signatures of this type of trap are 'sawtooth'-like diamonds, and abrupt shifts of ZBPs in external magnetic fields. Such features are not present in results reported here. Finally, perturbations of charge traps to the QD only happen occasionally, while the splittings are observed in every ZBP of every device in the work reported here.

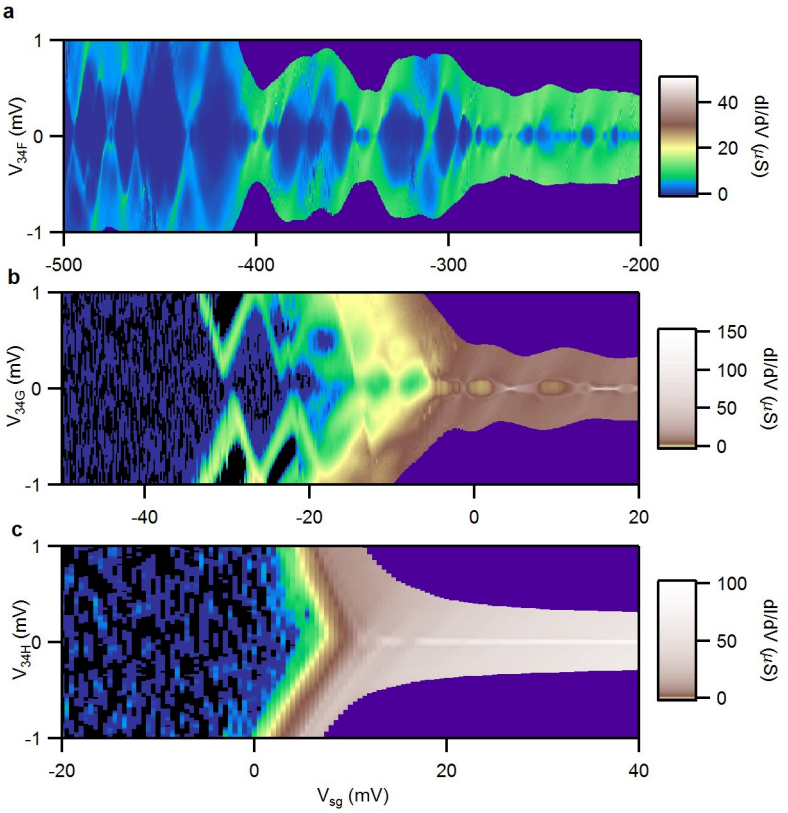
Sample size. No statistical methods were used to predetermine sample size.

31. Bi, F. *et al.* "Water-cycle" mechanism for writing and erasing nanostructures at the $LaAlO_3/SrTiO_3$ interface. *Appl. Phys. Lett.* **97,** 173110 (2010).
32. Levy, A. *et al.* Writing and low-temperature characterization of oxide nanostructures. *J. Vis. Exp.* **89,** e51886 (2014).
33. Irvin, P. *et al.* Rewritable nanoscale oxide photodetector. *Nature Photon.* **4,** 849–852 (2010).
34. Ma, Y. *et al.* Broadband terahertz generation and detection at 10 nm scale. *Nano Lett.* **13,** 2884–2888 (2013).
35. Ralph, D. C., Black, C. T. & Tinkham, M. Spectroscopic measurements of discrete electronic states in single metal particles. *Phys. Rev. Lett.* **74,** 3241–3244 (1995).
36. Beenakker, C. W. J. Theory of Coulomb-blockade oscillations in the conductance of a quantum dot. *Phys. Rev. B* **44,** 1646–1656 (1991).

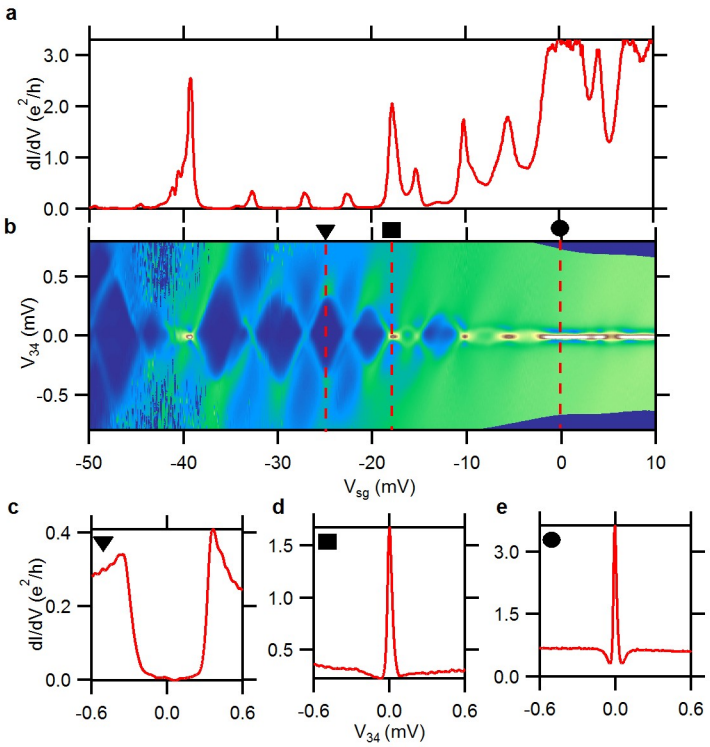
37. Müller, K. A. & Burkard, H. $SrTiO_3$ — Intrinsic quantum paraelectric below 4 K. *Phys. Rev. B* **19,** 3593–3602 (1979).
38. Santander-Syro, A. F. *et al.* Two-dimensional electron gas with universal subbands at the surface of $SrTiO_3$. *Nature* **469,** 189–193 (2011).
39. Averin, D. V. & Nazarov, Y. V. Single-electron charging of a superconducting island. *Phys. Rev. Lett.* **69,** 1993–1996 (1992).
40. Schlottmann, P. Exact results for highly correlated electron systems in one dimension. *Int. J. Mod. Phys. B* **11,** 355–667 (1997).
41. Lin, H. Q. Dilute gas of electron pairs in the t-J model. *Phys. Rev. B* **44,** 4674–4676 (1991).
42. Hellberg, C. S. & Manousakis, E. 2-dimensional t-J model at low electron density. *Phys. Rev. B* **52,** 4639–4642 (1995).
43. Cullum, J. & Willoughby, R. A. Computing eigenvalues of very large symmetric-matrices — an implementation of a Lanczos-algorithm with no reorthogonalization. *J. Comput. Phys.* **44,** 329–358 (1981).
44. Cullum, J. & Willoughby, R. A. A survey of Lanczos procedures for very large real symmetric eigenvalue problems. *J. Comput. Appl. Math.* **12–13,** 37–60 (1985).
45. Hellberg, C. S. in *Computer Simulation Studies in Condensed Matter Physics XIII* (eds Landau, D. P., Lewis, S. P. & Schüttler, H. B.) 43–52 (Springer, 2000).
46. Hellberg, C. S. Theory of the reentrant charge-order transition in the manganites. *J. Appl. Phys.* **89,** 6627–6629 (2001).
47. Anderson, P. W. Model for electronic structure of amorphous semiconductors. *Phys. Rev. Lett.* **34,** 953–955 (1975).
48. Ashoori, R. C. *et al.* Single-electron capacitance spectroscopy of discrete quantum levels. *Phys. Rev. Lett.* **68,** 3088–3091 (1992).
49. Alt, H. C. Experimental evidence for a negative-U center in gallium arsenide related to oxygen. *Phys. Rev. Lett.* **65,** 3421–3424 (1990).
50. Geballe, T. H. & Kivelson, S. A. Paired insulators and high temperature superconductors. Preprint at http://arXiv.org/abs/1406.3759 (2014).
51. Stashans, A., Pinto, H. & Sanchez, P. Superconductivity and Jahn-Teller polarons in titanates. *J. Low Temp. Phys.* **130,** 415–423 (2003).
52. Gervais, F., Servoin, J. L., Baratoff, A., Bednorz, J. G. & Binnig, G. Temperature dependence of plasmons in Nb-doped $SrTiO_3$. *Phys. Rev. B* **47,** 8187–8194 (1993).
53. van Mechelen, J. L. M. *et al.* Electron-phonon interaction and charge carrier mass enhancement in $SrTiO_3$. *Phys. Rev. Lett.* **100,** 226403 (2008).
54. Kolodiazhnyi, T. & Wimbush, S. C. Spin-singlet small bipolarons in Nb-doped $BaTiO_3$. *Phys. Rev. Lett.* **96,** 246404 (2006).
55. Jung, S. W., Fujisawa, T., Hirayama, Y. & Jeong, Y. H. Background charge fluctuation in a GaAs quantum dot device. *Appl. Phys. Lett.* **85,** 768–770 (2004).
56. Bolotin, K. I., Kuemmeth, F., Pasupathy, A. N. & Ralph, D. C. Metal-nanoparticle single-electron transistors fabricated using electromigration. *Appl. Phys. Lett.* **84,** 3154–3156 (2004).



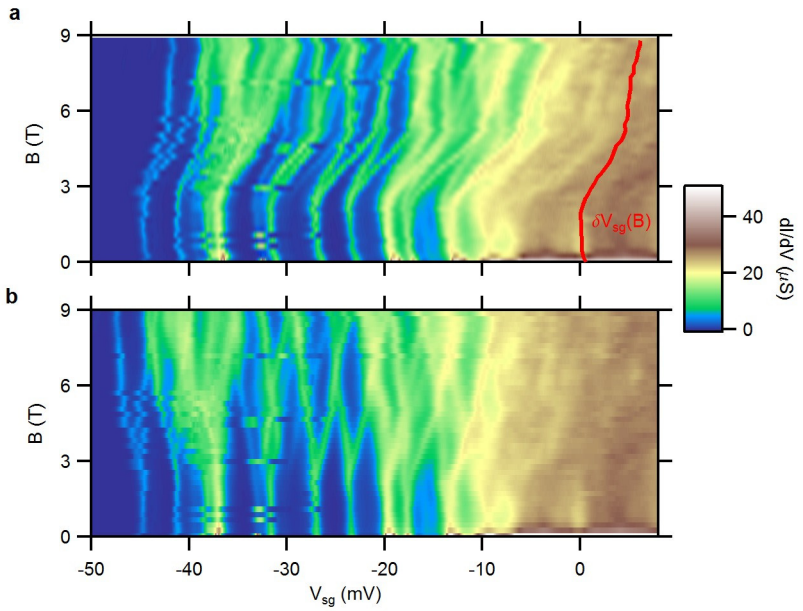
Extended Data Figure 1 | Nanoscale potential barrier engineering and low-temperature transport characteristics. **a**, Single-barrier device schematic. It has the same structure as device A except that only one barrier is integrated in the design. **b**, Resistance change during barrier cutting (Methods); t is time. **c**, The differential conductance $\mathrm{d}I/\mathrm{d}V$ as a function of V_{sg} at $T = 75$ mK. The wire can be pinched off by V_{sg} at the barrier site, as the wire conductance becomes negligibly small in lower V_{sg} values.



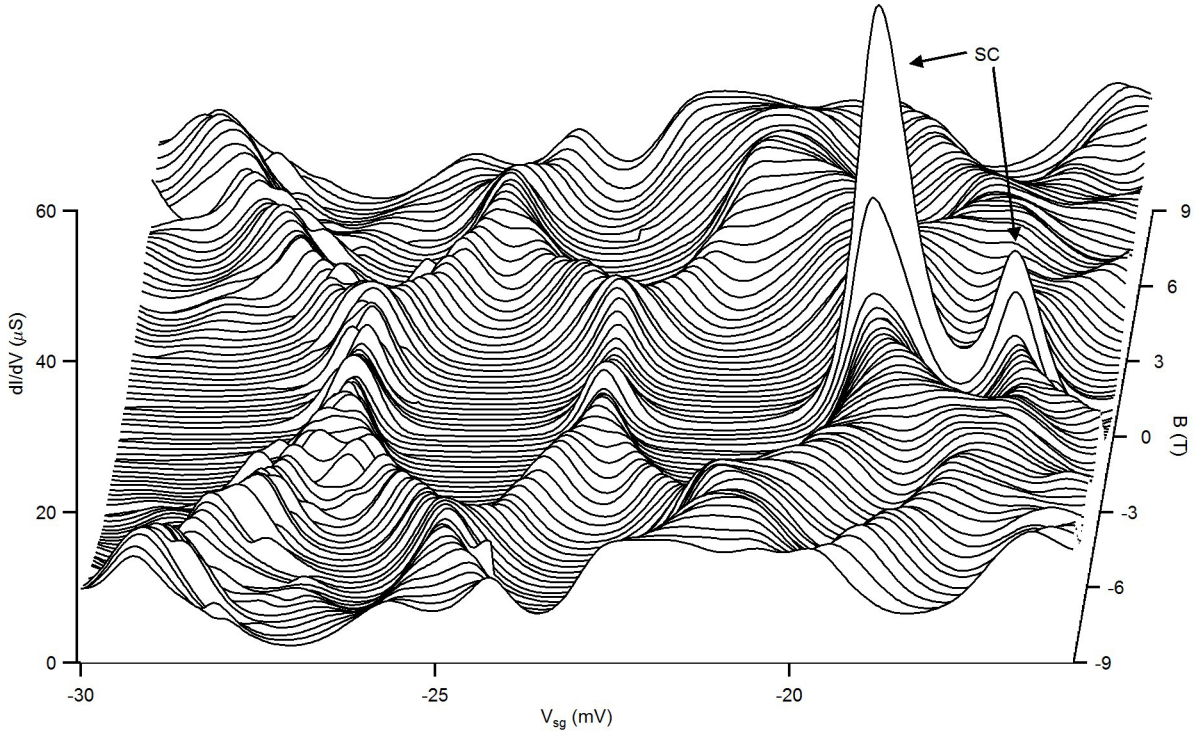
Extended Data Figure 2 | Transport properties of three L_{QD} = 500 nm SET devices of different single barrier heights at T = 50 mK. Plots show colour-coded dI/dV as a function of V_{sg} and voltage across the QDs in devices F (V_{34F}), G (V_{34G}) and H (V_{34H}). **a**, Device F ($\Delta R/2 = 20\,k\Omega$) requires a back gate voltage V_{bg} applied on the substrate of −5.6 V to pinch off the device since V_{sg} has limited tunability due to leakage at high absolute values. **b**, Device G ($\Delta R/2 = 110\,k\Omega$) shows similar properties to device A. **c**, Device H ($\Delta R/2 = 305\,k\Omega$) shows no conductance diamonds.



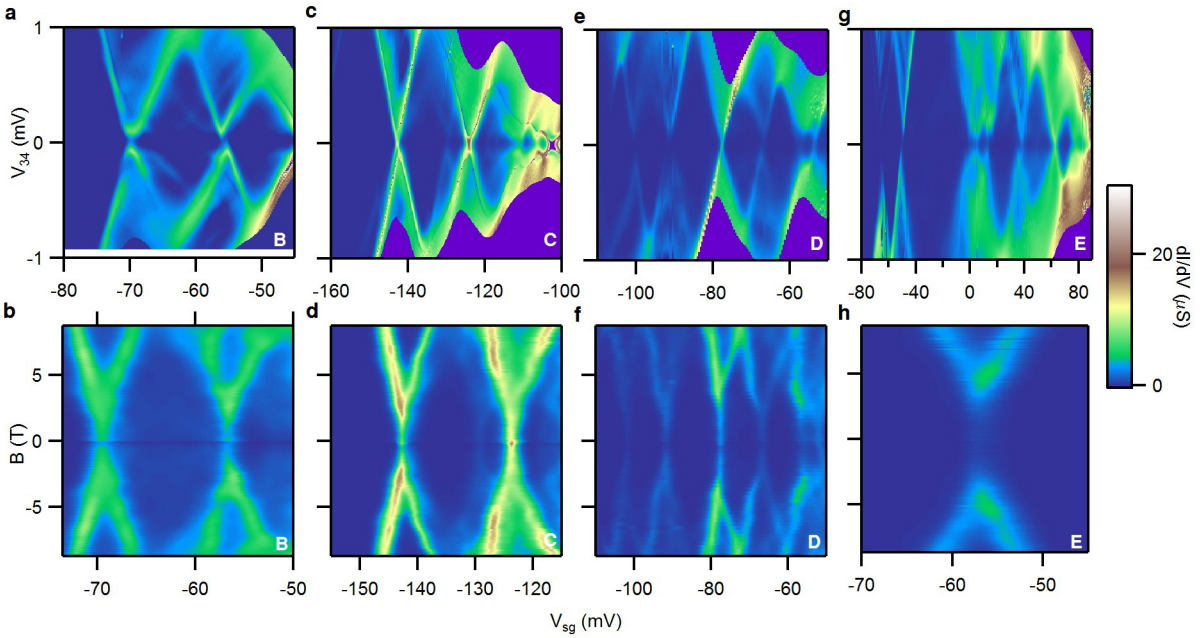
Extended Data Figure 3 | Transport characteristics due to barrier confinement of device A. **a**, Zero-bias line cut of the dI/dV map in **b**; filled black symbols show positions of line cuts displayed in **c**–**e**. **c**–**e**, Full suppression of transport in device A at $V_{sg} = -25$ mV (**c**), resonant tunnelling transport at $V_{sg} = -19$ mV (**d**), and fully superconducting transport at $V_{sg} = 0$ mV (**e**).



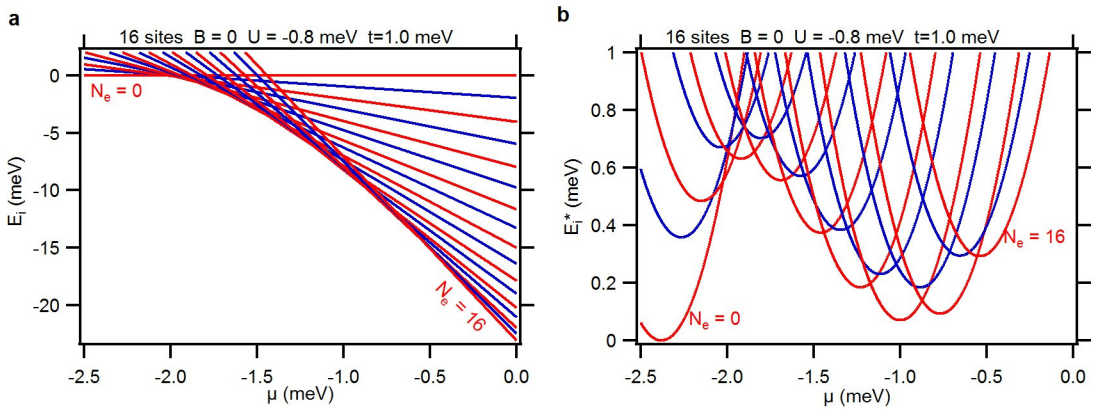
Extended Data Figure 4 | Global shift correction of data from device A. **a**, Original data for Figs 2f and 4a in the main text; *B*, applied magnetic field. The global shift is illustrated by the red trace, $\delta V_{sg}(B)$. **b**, The same data shown in Figs 2f and 4a in the main text that are corrected by $V_{sg}(B) = V_{sg0}(B) - \delta V_{sg}(B)$.



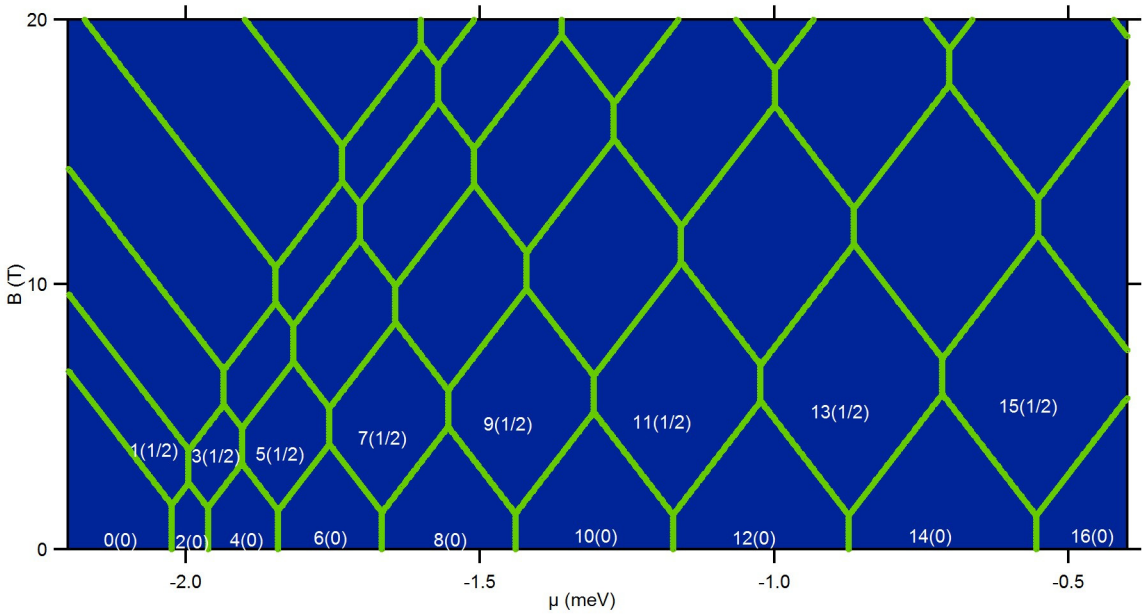
Extended Data Figure 5 | Three-dimensional 'waterfall' plot of the magnetic-field dependence of ZBPs (same data as shown in Fig. 2f). Plot shows lock-in dI/dV data at small (100 μV) bias as a function of V_{sg}, taken as the magnetic field is swept from −9 T to 9 T (additional right-axis).



Extended Data Figure 6 | Transport characteristics of devices B, C, D and E, which are all of the same geometry as device A. Device letter is shown at lower right-hand corner of all plots. **a**, Device B dI/dV (colour coded) as a function of V_{sg} and V_{34} at $V_{bg} = -0.7$ V and $T = 100$ mK. A small gap (4Δ) close to zero-bias in the diamonds is due to the absence of normal carriers in the superconducting source/drain leads. **b**, Device B ZBP splitting in an out-of-plane magnetic field B, which is the same data as Fig. 3a in the main text. **c**, Device C dI/dV as a function of V_{sg} and V_{34} at $V_{bg} = -4.4$ V and $T = 100$ mK. **d**, Device C ZBP splitting in an out-of-plane magnetic field. **e**, Device D dI/dV as a function of V_{sg} and V_{34} at $V_{bg} = -1.4$ V and $T = 100$ mK. **f**, Device D ZBP splitting in an out-of-plane magnetic field. **g**, Device E dI/dV dependent on V_{sg} and V_{34} at $V_{bg} = -2.2$ V and $T = 100$ mK. **h**, Device E ZBP splitting in an out-of-plane magnetic field.



Extended Data Figure 7 | Parity effect. **a**, Energies E_i of the Hubbard model (equation (1) in main text) of a one-dimensional 16-site chain with open boundary conditions, $t = 1$ meV, $U = -0.8$ meV, and $B = 0$ for fillings $N_e \leq 16$. The slope of each line is proportional to $-N_e$; red (blue) lines have even (odd) N_e. For all chemical potentials μ, the ground state has even N_e. **b**, Energies of the Hubbard model for the same parameters as **a** shifted by a quadratic function of μ, $E_i^*(\mu) = E_i(\mu) + c\mu^2$, where c is arbitrary. The lowest energy for each value of μ is easier to discern. The ground state always has even N_e.



Extended Data Figure 8 | Phase diagram of the Hubbard model on a one-dimensional 16-site chain with $t = 1$ meV and $U = -0.8$ meV. The total number of electrons N_e and total spin S are labelled for some of the larger phases as $N_e(S)$. The quantum numbers of the other phases can be deduced from their neighbours.

Extended Data Table 1 | Parameters of eight devices

Device Name	L_w (μm)	L_{QD} (μm)	ΔR / 2 (kΩ)	Coupling Factor Range (eV/V)
A	2.5	1	32	0.08-0.13
B	2.5	1	30	0.06-0.11
C	2.5	1	15	0.04-0.10
D	2.5	1	15	0.04-0.10
E	2.5	1	25	0.03-0.06
F	2	0.5	20	0.03-0.06
G	2	0.5	110	0.06-0.12
H	2	0.5	305	0.07-0.09

All devices have a schematic similar to that of device A but with different control (open) wire length L_w, distance between two barriers L_{QD}, single barrier resistance $\Delta R/2$, and range of side gate coupling factor α for all the diamonds.

LETTER

doi:10.1038/nature14463

Quantum coherent optical phase modulation in an ultrafast transmission electron microscope

Armin Feist[1], Katharina E. Echternkamp[1], Jakob Schauss[1], Sergey V. Yalunin[1], Sascha Schäfer[1] & Claus Ropers[1]

Coherent manipulation of quantum systems with light is expected to be a cornerstone of future information and communication technology, including quantum computation and cryptography[1]. The transfer of an optical phase onto a quantum wavefunction is a defining aspect of coherent interactions and forms the basis of quantum state preparation, synchronization and metrology. Light-phase-modulated electron states near atoms and molecules are essential for the techniques of attosecond science, including the generation of extreme-ultraviolet pulses and orbital tomography[2,3]. In contrast, the quantum-coherent phase-modulation of energetic free-electron beams has not been demonstrated, although it promises direct access to ultrafast imaging and spectroscopy with tailored electron pulses on the attosecond scale. Here we demonstrate the coherent quantum state manipulation of free-electron populations in an electron microscope beam. We employ the interaction of ultrashort electron pulses with optical near-fields[4–9] to induce Rabi oscillations in the populations of electron momentum states, observed as a function of the optical driving field. Excellent agreement with the scaling of an equal-Rabi multilevel quantum ladder is obtained[10], representing the observation of a light-driven 'quantum walk'[5] coherently reshaping electron density in momentum space[11]. We note that, after the interaction, the optically generated superposition of momentum states evolves into a train of attosecond electron pulses. Our results reveal the potential of quantum control for the precision structuring of electron densities, with possible applications ranging from ultrafast electron spectroscopy and microscopy to accelerator science and free-electron lasers.

The interaction of propagating light with confined electrons in atoms, molecules and solids is omnipresent, but the opposite case—the coupling of free electrons to localized optical fields—is not a naturally occurring phenomenon. Nonetheless, in both cases, the principle of confinement allows for optical transitions in otherwise mismatched electron and photon dispersion relations[12]. Controlling free-electron propagation with low-frequency electromagnetic fields in resonator geometries is an integral aspect of accelerator science[13]. At optical frequencies, however, particular challenges arise from the requirements of very controlled electron beams and tailored nanostructure near-fields. Increasing efforts are currently devoted to optically drive electron trajectories on the nanoscale—for example, for applications in attosecond science and lightwave electronics[14–19].

Some of the elementary phenomena involved in coupling free electrons to light were described more than half a century ago: in the Kapitza–Dirac effect[20,21], electrons are elastically scattered off a standing light wave, whereas the Smith–Purcell effect and its variants[8,9,22,23] treat the inelastic interaction of free electrons with confined modes close to a grating. Recently, ultrafast electron microscopy schemes showed that the kinetic energy distribution of short electron pulses develops a series of photon sidebands after passage through an intense optical near-field[4–6]. This approach, termed photon-induced near-field electron microscopy (PINEM)[4], has been employed in the temporal characterization of ultrashort electron pulses (see Methods) and as a contrast mechanism in electron microscopy[24,25]. Beyond such advanced applications, the underlying interaction should allow for the preparation of coherent electronic superposition states and a phase-controlled harnessing of quantum coherence for the temporal shaping of electron bunches.

Here we report the coherent phase-modulation of free-electron states in a nano-optical field. We experimentally induce multilevel Rabi oscillations in the form of a quantum walk in momentum space, obtaining excellent agreement with theoretical predictions by García de Abajo *et al.*[5] and Park *et al.*[6] of this interaction. Moreover, we demonstrate theoretically that dispersive propagation transforms the optically modulated electron wavepacket into a train of attosecond peaks. In the experimental scenario displayed in Fig. 1, femtosecond electron pulses are generated by nonlinear photoemission from a nanoscale cathode[26–28]. After collimation and acceleration to an energy of 120 keV, the magnetic lens system of a transmission electron microscope focuses the electron pulses to a spot diameter of 15 nm in close vicinity to an optically excited conical gold tip. The localization of the nanostructure's near-field mediates the optical interaction with the free electrons. This leads to the creation of multiple spectral sidebands, each corresponding to the absorption/emission of an integer number of photons (spectrum in Fig. 1e)[4–6]. Detailed information about the interaction process is encoded in the number of populated sidebands and their individual amplitudes. For example, the maximum electron energy gain in the optical near field is a quantitative measure of the local transition amplitude, which can be imaged by raster scanning the electron focus (Fig. 1b).

Microscopically, the electron–light interaction studied here constitutes an optical phase-modulation of the electron wavefunction[6]. Expressed as a quantum mechanical multilevel system, electron energy levels spaced by the photon energy $\hbar\omega$ are coupled in the optical near-field (level diagram, Fig. 1d). Previous experiments studying this interaction found a partial reduction of the initial electronic state population and a spectral broadening with distributions gradually decaying towards large photon orders[4,6,7,24,25]. Such observations evidence transitions dominated by sequential multilevel excitation (processes of type I in Fig. 1d). However, it is assumed that the coupling process is coherent in nature[5,6], which implies that quantum features arising from multipath interference (type II) should also be observable.

In order to identify such phenomena, we require an interaction of uniform strength with the entire electron ensemble in the pulse[5]. This scenario is achieved by using a spatially narrow probing beam and, in contrast to earlier works, an optical near-field excitation which has a uniform amplitude during the transit of the electron pulse envelope (see Methods).

Under these conditions, we find experimentally that the population of photon sidebands exhibits a pronounced oscillatory behaviour corresponding to multilevel Rabi oscillations, as demonstrated in Fig. 2 for electron spectra at a fixed position near the tip shaft. A colour-coded map (Fig. 2a) displays the evolution of the interaction-induced kinetic energy distribution with growing incident field strength. With increasing driving field, we observe a linear spreading in the range of

[1]4th Physical Institute, Solids and Nanostructures, University of Göttingen, Göttingen 37077, Germany.

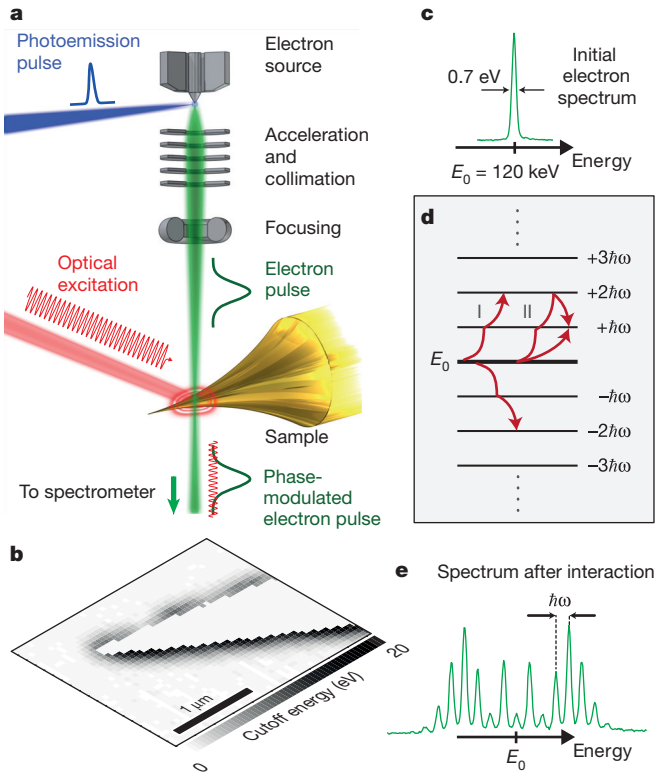


Figure 1 | Schematic and principles of coherent inelastic electron scattering by optical near-fields. **a**, Experimental scheme. Ultrashort electron pulses generated by nanotip photoemission are accelerated and focused to a beam that interacts with the optical near-field of a nanostructure, phase-modulating the electron pulse and exchanging energy in integer multiples of the photon energy. **b**, Raster-scanned image of the energy cutoff in the inelastic electron scattering spectra, representing the local transition amplitude (see text). **c**, Incident kinetic energy spectrum (full-width at half-maximum, 0.7 eV) centred at $E_0 = 120$ keV. **d**, Energy level diagram of ladder states with spacing $\hbar\omega$ coupled to the initial state at E_0. Arrows indicate sequential multistate population transfer (type I) and interfering quantum paths (type II) leading to multilevel Rabi oscillations. **e**, Example of kinetic energy spectrum after the near-field interaction, exhibiting a spectral comb with multiple sidebands separated by the photon energy and modulated in occupation.

populated sidebands, together with strong oscillations in the central part of the spectra. Specifically, the experimental spectra exhibit a nearly complete extinction of the initial state occupation and its pronounced recurrence at incident fields of 0.023 V nm^{-1} (red line) and 0.040 V nm^{-1} (green line), respectively (Fig. 2c). Quantitative analysis of the field-dependent spectral evolution (Fig. 2b) shows the oscillations of the initial state population ('zero loss peak') and those of different electronic sidebands. These modulations directly evidence multilevel Rabi oscillations and thereby a quantum coherent manipulation of the respective level amplitudes, which, as a function of field strength, traces out the evolution of an elementary quantum walk[11].

Recently, near-field-induced free-electron transitions, as observed here, were theoretically treated by solving the time-dependent Schrödinger equation[5,6] (compare our results to figure 2 in ref. 5, which depicts population oscillations simulated as a function of intensity). Yielding equivalent theoretical results, we present a compact description using raising and lowering operators acting on the electronic state $|E_0\rangle$ of the system at an initial energy E_0. As demonstrated in the Methods section, the action of the near-field can be described by a scattering matrix $S = \exp[g^* a - g a^\dagger]$ with a dimensionless near-field coupling constant g proportional to the field strength and the transition matrix element. One may notice that S takes the

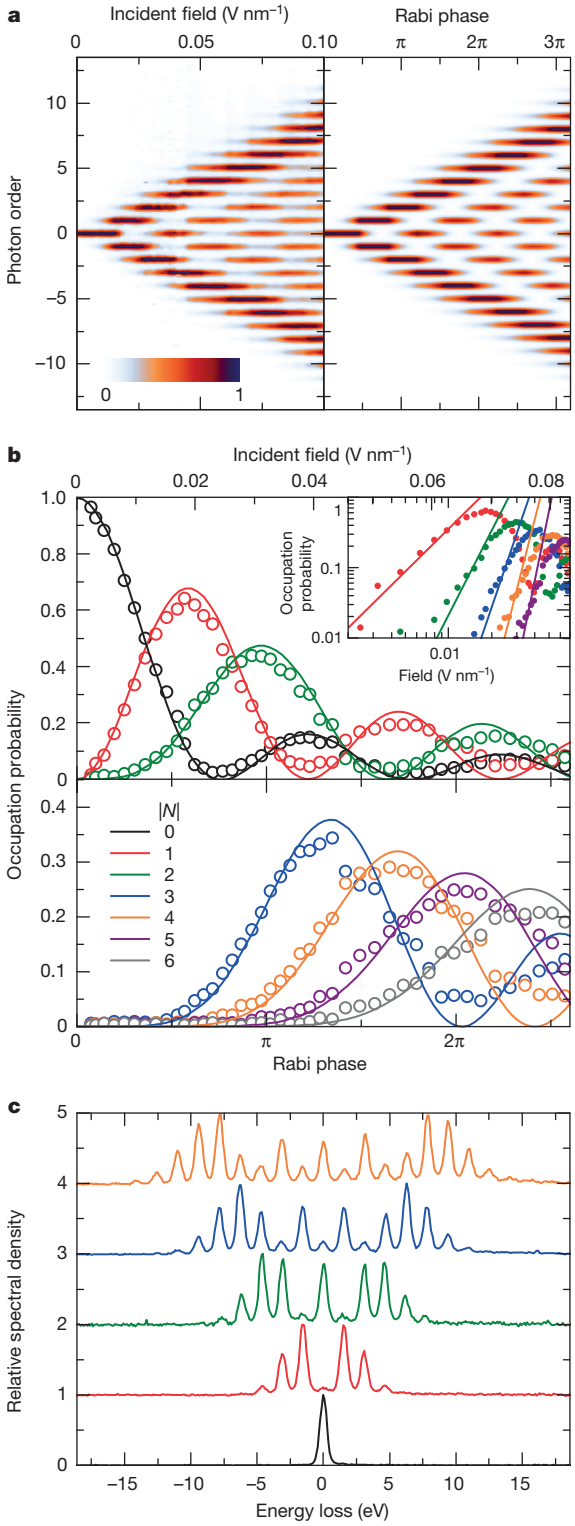


Figure 2 | Quantum coherent manipulation of electron energy distributions. **a**, Experimental electron energy distributions as a function of the incident optical field strength (left) and theoretical prediction in terms of Nth-order Bessel functions (right). **b**, Occupation probabilities (open circles) of the Nth-order spectral sidebands shown in **a**, adding contributions from $\pm|N|$. Solid lines, Nth-order Bessel functions. Inset, double-logarithmic plot of the sideband populations near onset. Solid lines with slope $2N$ are shown for comparison. **c**, Electron energy spectra at incident optical fields of 0, 0.023, 0.040, 0.053 and 0.068 V nm^{-1} (bottom to top). Spectra shown in **a** and **c** are normalized to their respective maxima.

form of a displacement operator generating coherent states in the harmonic oscillator model. In a variation of this scenario, we use $a^\dagger$ and a as commuting raising and lowering operators, connecting free-electron states separated by the photon energy, that is, $a^\dagger|E_0\rangle = |E_0 + \hbar\omega\rangle$ and $a|E_0\rangle = |E_0 - \hbar\omega\rangle$. The scattering matrix represents the Hamiltonian evolution of the system as a unitary operation on the initial electron wavefunction $|E_0\rangle$, which leaves the electron in a superposition of ladder states $|E_0 \pm N\hbar\omega\rangle$, with N a natural number.

In the limit of small optical driving fields, the scattering matrix is dominated by sequential multiphoton terms, for example, $\frac{1}{N!}\left(ga^\dagger\right)^N$, corresponding to type I transitions (Fig. 1d). The occurrence of interfering quantum paths at increased optical field strength (type II paths) becomes apparent by considering higher order terms in the Taylor expansion of S, such as $a^\dagger$ and $a^\dagger a^\dagger a$, which both facilitate the transition between the states $|E_0\rangle$ and $|E_0 + \hbar\omega\rangle$, but each with a different phase factor in the final state.

Interestingly, because of the practically constant coupling matrix elements between adjacent levels (the 'equal Rabi' case[10]), the occupation probability of the Nth photon sideband can be described by a very simple analytical expression in the form of the Nth-order Bessel function of the first kind[5,6], that is, $|\langle E_0 \pm N\hbar\omega|S|E_0\rangle|^2 = |J_N(2|g|)|^2$. In a spatial representation, these transitions arise from a sinusoidal phase modulation of the wavefunction traversing the optical near-field. Accordingly, such sideband populations are also commonly encountered in other physical systems using phase modulation, for example, in acousto-optics[29].

Comparing the experimental field-dependent electron populations with the analytical result (Fig. 2b), an excellent agreement is found both in the location and amplitudes of the respective occupation minima/maxima. The entire data set is described with a single Rabi phase $2|g| = F_{\mathrm{inc}} \times 98\,\mathrm{V}^{-1}$ nm linearly increasing with the incident optical field strength F_{inc}, yielding a quantitative measure of the transition matrix element. As detailed in the Methods section, incomplete modulation of the Rabi oscillations at higher fields is caused by the finite spatial and temporal electron pulse widths within the optical near-field. Besides the predicted population oscillations, the characteristic low-field multiphoton limit of the electron–light interaction is also experimentally regained (slopes of $2N$ in the field in a double-logarithmic plot, see inset of Fig. 2b). Larger incident fields prominently transfer the electron distribution to the outer spectral lobes, creating a well-defined cutoff around $|E - E_0| = 2|g|\hbar\omega$, equal to the maximum classical energy transfer. Thus, as in other instances of electrons driven by intense optical near-fields[15,16,19], the interaction energy is governed directly by the field amplitude instead of the ponderomotive energy.

This periodic phase (and correspondingly momentum) modulation of the electron wavefunction has important consequences for its subsequent evolution in free propagation. Generally, momentum modulation of classical states in particle accelerators is used for bunch compression[13], and an optical variant of this principle was recently proposed using ponderomotive forces acting on classical point particles[30]. However, the present conditions with electronic coherence times exceeding the optical period necessitate a quantum mechanical description of bunch reshaping. Figure 3a displays a few cycles of the simulated electron density in a periodically phase-modulated wavepacket as a function of the propagation distance and the arrival time at this distance relative to the centre of the pulse. Specifically, the phased superposition of momentum states reshapes into a high-contrast train of attosecond pulses at a well-defined distance downstream from the interaction region. For typical coupling constants achieved in the experiments, we obtain a temporal focusing into a train of pulses only about 80 as long, at a distance of 1.8 mm behind the sample. Further dispersion spreads the distribution corresponding to its momentum content, with the possibility of revivals. Note that a spatial optical equivalent of this generation of attosecond spikes is given by Fresnel diffraction at a sinusoidal phase grating into a near-field fringe pattern[31], and also that an early theoretical scheme for subfemtosecond optical pulse generation relied on frequency modulation and subsequent reshaping[32].

The physical origin of this electron pulse compression can be illustrated using a phase space representation of the quantum state in the form of a Wigner function. This function is a quantum mechanical analogue of a phase space density, which, however, can also take negative values for non-classical states[33]. Figure 3b displays the Wigner function of one period of a propagated state at the temporal focus and for a typical momentum distribution (projection in Fig. 3c). In this plane, free propagation of the initially sinusoidal momentum modulation has sheared the phase space distribution to a situation where a highly localized projection onto the position axis, that is, arrival time, is formed (Fig. 3d). In fact, the generation of this attosecond electron pulse train is very robust with respect to variations of the specific temporal and energetic structure of the initial electron pulse (see Methods). The practically usable focal distances (a few millimetres) render this scheme directly applicable in electron microscopy or spectroscopy studies with attosecond precision, a domain at present

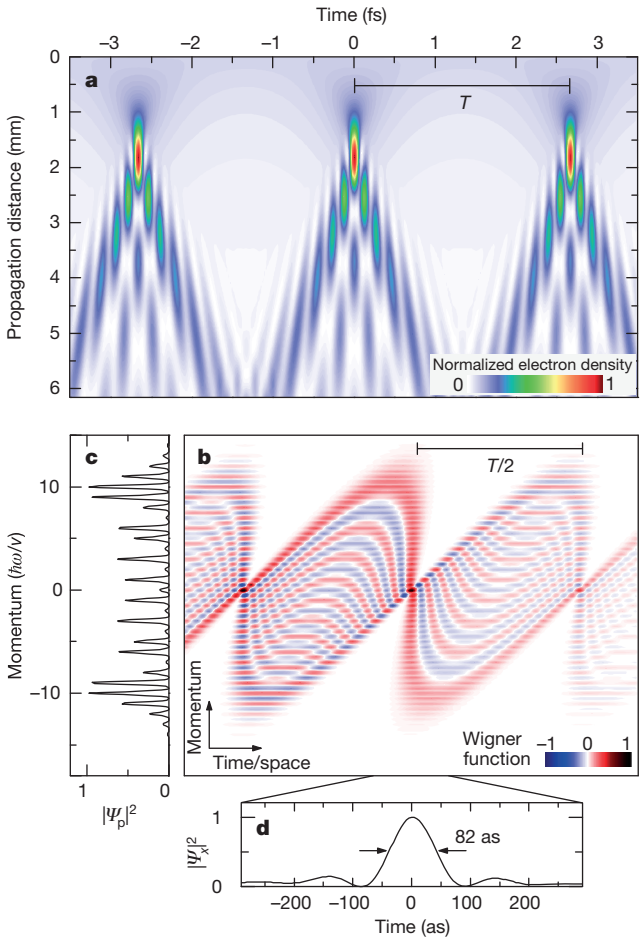


Figure 3 | Formation of an attosecond electron pulse train. **a**, Development of a periodically modulated electron pulse structure (normalized electron density is colour coded) as a function of the propagation distance after the near-field interaction (numerical simulation for $|g| = 5.7$). Free propagation causes a temporal focusing into a train of attosecond spikes (red) with a period of $T = 2.55$ fs (optical period). **b**, Phase space (Wigner) representation of one period of the light-modulated electron quantum state at the temporal focus position (propagation distance of 1.8 mm in **a**). Note that time and space variables for the swift electron pulse can be used equivalently via $x = vt$ (v is mean electron velocity). **c**, Momentum projection of Wigner function exhibiting spectral modulations as observed in the experiments, displayed in units of transferred momentum quanta (average momentum subtracted). **d**, Central part of spatial projection, expressed in terms of electron arrival time in laboratory frame. A peak with a duration of only 82 as (FWHM) is produced.

only accessible by attosecond light pulses[2]. Specifically, the temporal electron comb will enable the phase-resolved investigation of coherent sample excitations, thus tracing structural or electronic changes carrying optical phase information.

In conclusion, we have demonstrated the quantum coherent manipulation of free-electron wavefunctions by their interaction with nanoconfined light fields, observing near-perfect correspondence to the behaviour of a multilevel model Hamiltonian. Thinking beyond a single-variable state control, near-field interactions are expected to cause entanglement of longitudinal and transverse momentum components, and moreover, Coulomb interactions in a beam crossover will result in correlations between multiple electrons. Both features may be crucial for employing free electrons in quantum information technology[1]. Perhaps surprisingly, the generation of an attosecond electron train is a direct and natural consequence of this optical phase-modulation. We anticipate various applications of this concept in imaging and spectroscopy—for example, in the phase-resolved detection of coherent, resonantly driven polarizations in solid state materials—thus opening up the study of attosecond phenomena in electron microscopy.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 November 2014; accepted 24 March 2015.

1. Bouwmeester, D., Ekert, A. & Zeilinger, A. *The Physics of Quantum Information* (Springer, 2000).
2. Krausz, F. & Ivanov, M. Attosecond physics. *Rev. Mod. Phys.* **81,** 163–234 (2009).
3. Itatani, J. *et al.* Tomographic imaging of molecular orbitals. *Nature* **432,** 867–871 (2004).
4. Barwick, B., Flannigan, D. J. & Zewail, A. H. Photon-induced near-field electron microscopy. *Nature* **462,** 902–906 (2009).
5. García de Abajo, F. J., Asenjo-Garcia, A. & Kociak, M. Multiphoton absorption and emission by interaction of swift electrons with evanescent light fields. *Nano Lett.* **10,** 1859–1863 (2010).
6. Park, S. T., Lin, M. & Zewail, A. H. Photon-induced near-field electron microscopy (PINEM): theoretical and experimental. *New J. Phys.* **12,** 123028 (2010).
7. Kirchner, F. O., Gliserin, A., Krausz, F. & Baum, P. Laser streaking of free electrons at 25 keV. *Nature Photon.* **8,** 52–57 (2013).
8. Peralta, E. A. *et al.* Demonstration of electron acceleration in a laser-driven dielectric microstructure. *Nature* **503,** 91–94 (2013).
9. Breuer, J. & Hommelhoff, P. Laser-based acceleration of nonrelativistic electrons at a dielectric structure. *Phys. Rev. Lett.* **111,** 134803 (2013).
10. Shore, B. W. & Eberly, J. H. Analytic approximations in multi-level excitation theory. *Opt. Commun.* **24,** 83–88 (1978).
11. Bouwmeester, D., Marzoli, I., Karman, G., Schleich, W. & Woerdman, J. Optical Galton board. *Phys. Rev. A* **61,** 013410 (1999).
12. García de Abajo, F. J. Optical excitations in electron microscopy. *Rev. Mod. Phys.* **82,** 209–275 (2010).
13. Hemsing, E., Stupakov, G., Xiang, D. & Zholents, A. Beam by design: laser manipulation of electrons in modern accelerators. *Rev. Mod. Phys.* **86,** 897–941 (2014).
14. Krüger, M., Schenk, M. & Hommelhoff, P. Attosecond control of electrons emitted from a nanoscale metal tip. *Nature* **475,** 78–81 (2011).
15. Herink, G., Solli, D. R., Gulde, M. & Ropers, C. Field-driven photoemission from nanostructures quenches the quiver motion. *Nature* **483,** 190–193 (2012).
16. Wimmer, L. *et al.* Terahertz control of nanotip photoemission. *Nature Phys.* **10,** 432–436 (2014).
17. Schiffrin, A. *et al.* Optical-field-induced current in dielectrics. *Nature* **493,** 70–74 (2012).
18. Piglosiewicz, B. *et al.* Carrier-envelope phase effects on the strong-field photoemission of electrons from metallic nanostructures. *Nature Photon.* **8,** 37–42 (2013).
19. Stockman, M. I., Kling, M. F., Kleineberg, U. & Krausz, F. Attosecond nanoplasmonic-field microscope. *Nature Photon.* **1,** 539–544 (2007).
20. Kapitza, P. L. & Dirac, P. M. The reflection of electrons from standing light waves. *Math. Proc. Camb. Phil. Soc.* **29,** 297–300 (1933).
21. Freimund, D. L., Aflatooni, K. & Batelaan, H. Observation of the Kapitza-Dirac effect. *Nature* **413,** 142–143 (2001).
22. Smith, S. & Purcell, E. Visible light from localized surface charges moving across a grating. *Phys. Rev.* **92,** 1069 (1953).
23. Mizuno, K., Pae, J., Nozokido, T. & Furuya, K. Experimental evidence of the inverse Smith-Purcell effect. *Nature* **328,** 45–47 (1987).
24. Flannigan, D. J., Barwick, B. & Zewail, A. H. Biological imaging with 4D ultrafast electron microscopy. *Proc. Natl Acad. Sci. USA* **107,** 9933–9937 (2010).
25. Yurtsever, A., van der Veen, R. M. & Zewail, A. H. Subparticle ultrafast spectrum imaging in 4D electron microscopy. *Science* **335,** 59–64 (2012).
26. Ropers, C., Solli, D. R., Schulz, C. P., Lienau, C. & Elsaesser, T. Localized multiphoton emission of femtosecond electron pulses from metal nanotips. *Phys. Rev. Lett.* **98,** 043907 (2007).
27. Hommelhoff, P., Kealhofer, C. & Kasevich, M. A. Ultrafast electron pulses from a tungsten tip triggered by low-power femtosecond laser pulses. *Phys. Rev. Lett.* **97,** 247402 (2006).
28. Gulde, M. *et al.* Ultrafast low-energy electron diffraction in transmission resolves polymer/graphene superstructure dynamics. *Science* **345,** 200–204 (2014).
29. Moharam, M. G. & Young, L. Criterion for Bragg and Raman-Nath diffraction regimes. *Appl. Opt.* **17,** 1757–1759 (1978).
30. Baum, P. & Zewail, A. H. Attosecond electron pulses for 4D diffraction and microscopy. *Proc. Natl Acad. Sci. USA* **104,** 18409–18414 (2007).
31. Case, W. B., Tomandl, M., Deachapunya, S. & Arndt, M. Realization of optical carpets in the Talbot and Talbot-Lau configurations. *Opt. Express* **17,** 20966–20974 (2009).
32. Harris, S. E. & Sokolov, A. V. Subfemtosecond pulse generation by molecular modulation. *Phys. Rev. Lett.* **81,** 2894–2897 (1998).
33. Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, 1995).

Acknowledgements We thank M. Sivis and B. Schröder for help with sample preparation. We also thank our colleagues within the Göttingen UTEM initiative (C. Jooß, M. Münzenberg, K. Samwer, M. Seibt, C.A. Volkert). This work was supported by the Deutsche Forschungsgemeinschaft (DFG-SFB 1073/project A05), the VolkswagenStiftung, and the Lower Saxony Ministry of Science and Culture. We thank JEOL Ltd and JEOL Germany for their continuing support during the development of the Göttingen Ultrafast Transmission Electron Microscope.

Author Contributions The experiments were carried out by A.F., with contributions from J.S. and S.S.; S.S. and C.R. conceived and directed the study; S.V.Y. developed the analytical description and K.E.E. carried out the numerical simulations, each with contributions from A.F., S.S. and C.R.; the manuscript was written by A.F., K.E.E., S.S. and C.R., after discussions with and input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S. (schaefer@ph4.physik.uni-goettingen.de) or C.R. (cropers@gwdg.de).

METHODS

Ultrafast TEM and experimental setup. We have recently constructed an ultrafast TEM (UTEM) to enable a variety of laser-pump/electron-probe imaging schemes with high spatial resolution. The microscope is based on a commercial Schottky field emission TEM (JEOL JEM-2100F), which we modified to allow for both optical sample excitation and laser-driven electron pulse generation in the gun, as shown in Extended Data Fig. 1. In contrast to previous implementations of time-resolved TEM, our instrument features a nanotip photocathode as the source of ultrashort electron pulses. Compared to planar emitters, such needle cathodes provide reduced electron beam emittance, which is particularly useful for nanoscale probing and spectroscopy. In the present experiments, we employ electron pulses with a repetition rate of 250 kHz, which are induced via two-photon photoemission by irradiating the apex of a tungsten field emission cathode (apex radius of curvature of about 120 nm) with ultrashort laser pulses (400 nm central wavelength, 50 fs pulse duration, 4.1 nJ pulse energy, 25 GW cm^{-2} incident peak intensity). The emitter tip is operated at room temperature and with an electrostatic extraction field of 0.1 V nm^{-1}.

At an electron energy of 120 keV, the beam is focused to a spot diameter of about 15 nm, with a typical electron pulse duration of 700–900 fs (full-width at half-maximum, FWHM) at the position of the sample (characterized by electron–laser cross-correlation, see below). This pulse duration is at present governed by velocity dispersion of electrons with different initial kinetic energies after photoemission, which, however, is not a limitation for the measurements described here. While pulses may contain many electrons, all experiments reported here were acquired with less than one electron per pulse at the sample position (single-electron mode), thus avoiding potential space charge influences.

The laser pulse exciting the optical near-field (central wavelength of 800 nm, pulse energy of up to 60 nJ) is stretched by dispersion in glass to a pulse duration of 3.4 ps and focused to a spot size of ~50 μm, yielding peak intensities of 1.33 GW cm^{-2} at the largest applied fluence. The excitation with a laser pulse of a duration much longer than that of the electron pulse allows for the observation of coherent population oscillations (see section 'Numerical calculations' below).

After interaction with the optical near-field, the electron pulse is imaged (magnification of 25,000) into an electron energy-loss spectrometer (EELS) to analyse its kinetic energy distribution (spectrometer entrance aperture of 3 mm, energy dispersion of 0.05 eV per detector channel).

Modulated electron spectra are observed at any electron probe position within the optical near-field. However, the quality of the spectral modulation and full extinction of individual orders crucially depends on the transverse homogeneity of the near-field on the scale of the electron beam diameter (see section 'Numerical calculations' below). For a nanoscopic tip, the optical near-fields are only slowly varying along its shaft. Therefore, we chose a position of the electron focus several micrometres away from the tip apex (see Extended Data Fig. 2).

Temporal characterization of electron pulses. In UTEM, the duration of the electron pulses is governed mainly by Coulomb repulsion and dispersive broadening within the electron gun and subsequent electron optics. The resulting electron pulse structure in the sample plane can be quantitatively characterized employing inelastic near-field electron scattering[34,35,42], as discussed in the main text. However, to this end, and in contrast to the experimental conditions described in the previous section, laser pulses of a duration (here 50 fs) much shorter than the electron pulse duration should be used in an electron/near-field cross-correlation[35].

Extended Data Fig. 3a displays electron energy spectra as a function of the temporal delay between the optical excitation and the electron pulse arrival, subtracting the electron spectrum in the absence of near-field excitation. Therefore, the central blue feature corresponds to the reduction of the zero-loss peak due to scattering of electrons into multiple photon sidebands (red stripes). The width of these features (Extended Data Fig. 3b) and their tilt in the energy–time diagram (inset) provide measures of the electron pulse duration and chirp, respectively. Here, we obtain a cross-correlation with a FWHM of 800 fs (standard deviation of 340 fs) and a chirp of ~760 fs eV^{-1} for an initial energy spread of 1.3 eV. In the experiments presented in the main text, electron pulses with a narrower energy spread of 0.7 eV were achieved. For the much shorter near-field excitation employed here, the scattering signal is linear in the momentary density of electrons within the pulse. Therefore, we can extract an energy–time representation of the electron pulse by superimposing all tilted sidebands (Extended Data Fig. 3c). Finally, we emphasize that temporally stretching the near-field excitation to more than 3 ps in the experiments ensures that it lasts much longer than the electron pulse duration and, therefore, provides for a nearly homogeneous scattering amplitude throughout the electron pulse.

Data analysis and drift correction. In order to obtain high quality electron spectra at a fixed sample location for varying driving field strengths, it has to be ensured that the laser-induced sample displacement, for example, by thermal expansion, is compensated for. This was achieved by first characterizing the fluence-dependent sample shift in imaging mode (observed up to 150 nm) and by automated electron beam repositioning between experimental runs. In addition, slow residual drifts (up to 20 nm, see Extended Data Fig. 2a) were corrected for by continuous line scans perpendicular to the gold surface and using the strength of the bulk plasmon band to identify the beam–surface distance. In the recorded spectra, energy losses due to bulk plasmon excitation generate a weak and spectrally broad band at energies above 15 eV (ref. 36), which is only present when the electron beam is placed in close proximity to the tip so that the outer tail of the electron focal spot grazes the tip surface. The plasmon contribution is well-separated from the main spectral features and can be easily subtracted from the spectra. Specifically, we identify the sideband populations by adopting a global fit function containing pseudo-Voigt profiles $V_p(E)$ for the zero-loss peak and all photon sidebands (using symmetric amplitudes in $\pm|N|$). The plasmon peak at a loss energy E_{pl} was described by an asymmetric Gaussian $G_{pl}(E)$:

$$P(E) = G_{pl}(E - E_{pl}) + \sum_{N=-\infty}^{\infty} a_{|N|} V_p(E + N\hbar\omega) \quad (1)$$

Extended Data Fig. 4 shows a typical electron spectrum together with the fitted function. Note that an energetic shift of the sideband comb relative to the zero-loss peak due to electron chirp is absent in the case of a long excitation pulse relative to the electron pulse duration and does not have to be included in the fit. An evaluation of the strength of the plasmon band as a function of beam–surface distance allows for a positioning accuracy of ±5 nm.

Materials. The nanostructure employed in this work was prepared from a thin gold wire (diameter 250 μm) which was subjected to thermal annealing in vacuum (800 °C, 12 h) to increase crystallinity and reduce surface roughness[37]. A sharp tip (100 nm apex radius) was formed by electrochemical etching in aqueous hydrochloric acid (37%)[38]. Afterwards, the conical part of the tip (length ~50 μm) was cut by focused ion beam milling, transferred to a silicon frame and attached by ion-beam deposited platinum.

Quantum description using ladder operators. The interaction of electrons traversing an optical near-field has been theoretically treated several times in the past, usually by either direct integration of the time-dependent Schrödinger equation[6] or using a Green's function formalism[5,39]. The relation of this stimulated process to the spontaneous mechanisms observed by electron energy-loss spectroscopy and cathodoluminescence is discussed in ref. 39. Furthermore, it was also shown that a non-relativistic approach is sufficient as long as the relativistically correct electron dispersion (velocity as a function of energy) is used in the final result[40]. Here, we present an alternative derivation of inelastic near-field scattering probabilities using ladder operators, which allows for a succinct description.

The raising and lowering operators. Electrons in a time-harmonic electromagnetic field can experience energy loss or gain in multiples of the photon energy $\hbar\omega$, where ω is the frequency of the field. This allows us to treat the problem as a multilevel quantum system. Within the Schrödinger picture, the free-electron Hamiltonian H_0 does not depend on time, while the wavefunction $|\psi(t)\rangle$ of the electron is time-dependent. Thus, the total Hamiltonian for the interaction with the electromagnetic field in the velocity gauge is

$$H = H_0 + \frac{e}{m} pA \quad (2)$$

where A is the space- and time-dependent vector potential, and p, e and m are the electron momentum, charge and mass, respectively. For a time-harmonic vector potential, a natural basis set is composed of plane wave states $|N\rangle$ offset from the initial energy E_0 by an integer multiple N of the photon energy, where each state $|N\rangle$ is an eigenstate of the unperturbed Hamiltonian H_0: $H_0|N\rangle = (E_0 + N\hbar\omega)|N\rangle$. Thus, $|0\rangle$ is the initial state, and $|N\rangle$ corresponds to the state with $|N|$ absorbed/emitted quanta. The time-harmonic interaction Hamiltonian causes transitions between these basis states. In particular, the matrix elements between adjacent states of the form $\langle N+1|\frac{e}{m}pA|N\rangle$ will lead to considerable transition probabilities. In contrast, the coupling between states separated by more than one photon energy causes transition amplitudes rapidly oscillating in time (at multiple frequencies of ω), which prevents direct multiphoton transitions. (Note that multiphoton transitions will become possible by multiple actions of the field.)

In order to compute the coupling between neighbouring states, let us consider for simplicity a one-dimensional model with the time-harmonic vector potential $A = F(z)\sin(\omega t)/\omega$, where $F(z)$ is the spatial distribution of the electric field amplitude. To obtain the matrix elements $\langle N+1|\frac{e}{m}pA|N\rangle$, we use $|N\rangle$ in a plane wave form $L^{-\frac{1}{2}}\exp(ik_N z)$ in a finite spatial interval of length L, where $\hbar k_N$ is the electron

momentum. In this representation, the matrix elements can be readily computed, for instance

$$\left\langle N+1\left|\frac{e}{m}pA\right|N\right\rangle=\frac{2\hbar v_N g}{L}\sin(\omega t),\quad g=\frac{e}{2\hbar\omega}\int_{-L/2}^{L/2}F(z)\exp(-i\Delta kz)\mathrm{d}z \tag{3}$$

where v_N is the electron velocity in the state $|N\rangle$, and $\Delta k\approx\omega/v_N$ is the electron momentum change (divided by $\hbar$). The dimensionless coupling constant g expressed in terms of a Fourier amplitude in equation (3) was introduced in a similar form as used in ref. 6. Physically, g describes the momentum component in the near-field distribution which allows for total energy and momentum conservation in the transition, that is, it represents the momentum change of an electron undergoing an energy transition of $\hbar\omega$. Regarding the integration limits in equation (3), at present, it is only important that the interval length L is larger than the extension of the near-field, as L will cancel out in the final result. It should be noted that for an initial energy much higher than the maximum number of absorbed or emitted photons, $E_0\gg|N|\hbar\omega$, the coupling matrix elements in equation (3) become practically independent of N, as does the velocity $v\approx v_N$. The presence of a single and universal coupling constant renders the present quantum system a nearly perfect example of an equal Rabi multilevel system[10]. The transitions in this system can be concisely described by introducing the raising and lowering operators $a^\dagger$ and a, respectively, as

$$|N+1\rangle=a^\dagger|N\rangle,\quad |N-1\rangle=a|N\rangle \tag{4}$$

Note that, in contrast to the commonly employed ladder operators of a harmonic oscillator (which has a coupling constant scaling with $\sqrt{N}$), it follows from equation (4) that a and $a^\dagger$ commute: $aa^\dagger=a^\dagger a$. The essential parts of the interaction Hamiltonian then take a bi-diagonal form, which can be represented in the raising and lowering operators

$$\frac{e}{m}pA=\frac{2\hbar v}{L}\left(g^*a+ga^\dagger\right)\sin(\omega t)+\mathcal{O}\left(a^n,a^{\dagger n};n\geq 2\right) \tag{5}$$

The higher order contributions can be neglected in the following, as they lead to negligible transition probabilities (see below), and terms on the main diagonal are absent because the spatial integral over the near-field distribution $F(z)$ (the case of $\Delta k=0$) vanishes.

The S-matrix. To obtain transition probabilities for electrons after passage through the near-field, it is convenient to switch to the interaction picture. Here, the lowering and raising operators become time-dependent: $a(t)$, $a^\dagger(t)$. In our case, they can be easily expressed in terms of a and $a^\dagger$ by the transformation

$$a(t)=\exp(-i\omega t)a,\quad a^\dagger(t)=\exp(i\omega t)a^\dagger \tag{6}$$

and the interaction Hamiltonian turns into

$$H_{\mathrm{int}}(t)=\frac{2\hbar v}{L}\sin(\omega t)\left[\exp(-i\omega t)g^*a+\exp(i\omega t)ga^\dagger\right] \tag{7}$$

where a and $a^\dagger$ denote the time-independent lowering and raising operators (see equation (4)). The temporal evolution of the quantum system can be treated in terms of a scattering matrix S, defined as a unitary transformation connecting asymptotic particle states $|\psi(\infty)\rangle=S|\psi(-\infty)\rangle$ before and after the interaction (for the time-dependence of the electron wavefunction during near-field transit, see ref. 5). This unitary operator S is given by the time-ordered exponent

$$S=T\exp\left(-\frac{1}{\hbar}\int_{-\infty}^{\infty}H_{\mathrm{int}}(t)\mathrm{d}t\right) \tag{8}$$

In the present case, the time-ordering T can be omitted because $a(t)$ and $a^\dagger(t)$ commute. With the choice of a finite support L of the basis states, the range of integration should in principle be limited to $\int_{-L/2v}^{L/2v}$, which will cancel out the ratio v/L appearing in equation (7) for the time-independent contributions. The terms oscillating at higher frequencies ($\int\exp(2i\omega t)\mathrm{d}t$ in equation (7) and higher order contributions from equation (5)) vanish in the limit of large L. This case of large L ($L>v/\omega$) corresponds to the experimental situation, in which the momentum states are well-resolved with respect to their energy difference $\hbar\omega$, and therefore, the passage to infinity can be carried out without loss of generality:

$$\frac{1}{\hbar}\int_{-\infty}^{\infty}H_{\mathrm{int}}(t)\mathrm{d}t=\frac{g^*a-ga^\dagger}{i} \tag{9}$$

Thus, the S-matrix in the interaction picture can be finally written as

$$S=\mathrm{e}^{ga^\dagger-g^*a} \tag{10}$$

and interestingly, the scattering matrix takes on the form of a displacement operator[33].

The transition probabilities. Using the S-matrix, we can compute the probabilities of the transitions $|0\rangle\rightarrow|N\rangle$, given by $P_N=|\langle N|S|0\rangle|^2$. For this purpose, we first split the matrix exponent in equation (10) into a product of two exponents, $\exp(ga^\dagger-g^*a)=\exp(ga^\dagger)\exp(-g^*a)$. This separation is of course possible because a and $a^\dagger$ commute. Expanding the exponential operators in a Taylor series, we find

$$\exp(ga^\dagger)|0\rangle=\sum_{m=0}^{\infty}\frac{g^m}{m!}(a^\dagger)^m|0\rangle=\sum_{m=0}^{\infty}\frac{g^m}{m!}|m\rangle \tag{11}$$

and analogously

$$\langle N|\exp(-g^*a)=\sum_{n=0}^{\infty}\frac{(-g^*)^n}{n!}\langle n|a^n=\sum_{n=0}^{\infty}\frac{(-g^*)^n}{n!}\langle N+n| \tag{12}$$

Using equation (10) and the orthogonality relation $\langle N+n|m\rangle=\delta_{N+n,m}$, we obtain

$$\langle N|S|0\rangle=\sum_{n=0}^{\infty}\frac{(-g^*)^n g^{n+N}}{n!(n+N)!}=g^N\sum_{n=0}^{\infty}\frac{\left(-|g|^2\right)^n}{n!(n+N)!} \tag{13}$$

Comparing this result with the following series expansion for the Bessel function of the first kind

$$J_N(z)=(z/2)^N\sum_{n=0}^{\infty}\frac{\left(-z^2/4\right)^n}{n!(n+N)!} \tag{14}$$

we finally obtain

$$P_N=J_N(2|g|)^2 \tag{15}$$

Therefore, the probability of energy gain or loss is given in the form of Bessel functions of different order[6].

Propagation after interaction and Wigner function. The propagation of the electron wavefunction after interaction with the optical near-field can be described in terms of a unitary evolution operator $\exp(-iH_0t/\hbar)$, where H_0 is again the free-electron Hamiltonian. Let $\psi_p(t)$ be the wavefunction in momentum representation and p the electron momentum in the laboratory frame. The unitary evolution is then given by

$$\psi_p(t)=\mathrm{e}^{-iE_pt/\hbar}\psi_p(0) \tag{16}$$

where $E_p=c\sqrt{(mc)^2+p^2}$ is the relativistic energy and m is the electron rest mass. In practice, the electron momenta p after the interaction are all very close to the initial (relativistic) electron momentum γmv, where v and γ are the initial electron velocity and the Lorentz contraction factor, respectively. For that reason, it is convenient to use 'shifted' momenta defined as $p'=p-\gamma mv$.

During the free propagation, the momentum distribution $|\psi_p(t)|^2$ remains unchanged because the unitary action only changes phases of the probability amplitudes $\psi_p(0)$. In contrast to the momentum distribution, the spatial density distribution will vary in time during the propagation. In a 'shifted' laboratory frame, the spatial representation of the wavefunction is given by the Fourier transformation

$$\psi(z-vt,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{ipz/\hbar}\psi_p(t)\mathrm{d}p \tag{17}$$

where $\psi_p(t)$ is normalized to unity and $z-vt$ is the shifted spatial coordinate. Figure 3a in the main text presents a computation of the probability density versus the arrival time of the wavepacket in a given plane as a function of the propagation distance between the interaction region and this plane.

The Wigner function of the quantum state (Fig. 3b in the main text) is given by

$$W(z-vt,p,t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\psi_{p+q}(t)\psi^*_{p-q}(t)\mathrm{e}^{-2iqz/\hbar}\mathrm{d}q \tag{18}$$

It gives a phase space representation of the quantum state[33] and illustrates the propagation-induced pulse compression.

Numerical calculations. In order to elucidate the importance of a spatially narrow probing beam and a temporally stretched near-field excitation for the observation of multilevel Rabi oscillations, we performed numerical calculations to quantitatively characterize the influence of an incoherent averaging over (temporally and spatially) varying transition probabilities $P_N(\boldsymbol{r}_\perp,t)$, where t is the electron arrival time and $\boldsymbol{r}_\perp$ its position vector in the sample plane (perpendicular to the beam

direction). In a different context, that is, in the description of laser-electron cross-correlations, similar computations were carried out in refs 5 and 6.

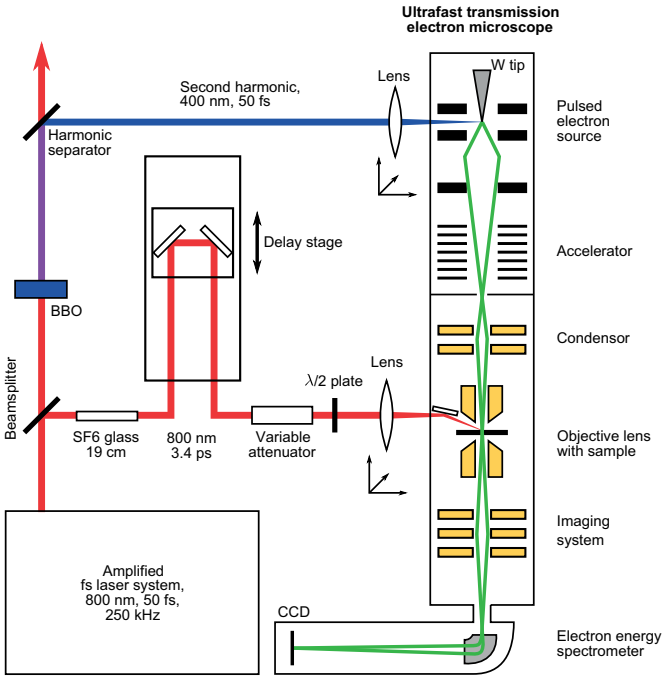
In a simplified geometry, the nanotip is modelled as a straight cylinder of a radius corresponding to that of the tip at the probing position ($r = 1\,\mu m$), for which the scattered electric field can be analytically calculated within Mie theory[41]. The field enhancement at the surface of the tip shaft is about 1.4. In the approximated geometry, we obtain a Fourier amplitude of the scattered field of $\frac{g}{F_{inc}} \approx 80.3\,V^{-1}nm$ (normalized to the incident field F_{inc}), in the same order of magnitude as in the experiments, and exponentially depending on the distance to the surface with a radial decay length of approximately 90 nm.

In Extended Data Fig. 5a, we study the effect of spatial and temporal averaging on the visibility of the Rabi oscillations by averaging over a disk-shaped beam and a Gaussian temporal distribution of different widths. The upper and lower graphs show the influence of a finite probing area and a reduction of the duration of sample excitation, respectively. Both for larger electron beam widths as well as shorter laser pulses driving the optical near-field, the Rabi oscillations exhibit weaker modulation and become substantially damped. Therefore, for the experimental electron pulse width of about 800 fs and a field decay of 90 nm, a probe radius around 10 nm and a near-field duration of 3.4 ps as in the experiments (black lines) allow for the observation of strongly modulated Rabi oscillations. For these experimental parameters, the sideband populations closely follow the analytical Bessel function dependence with minor deviations at higher fields (compare Extended Data Fig. 5b).

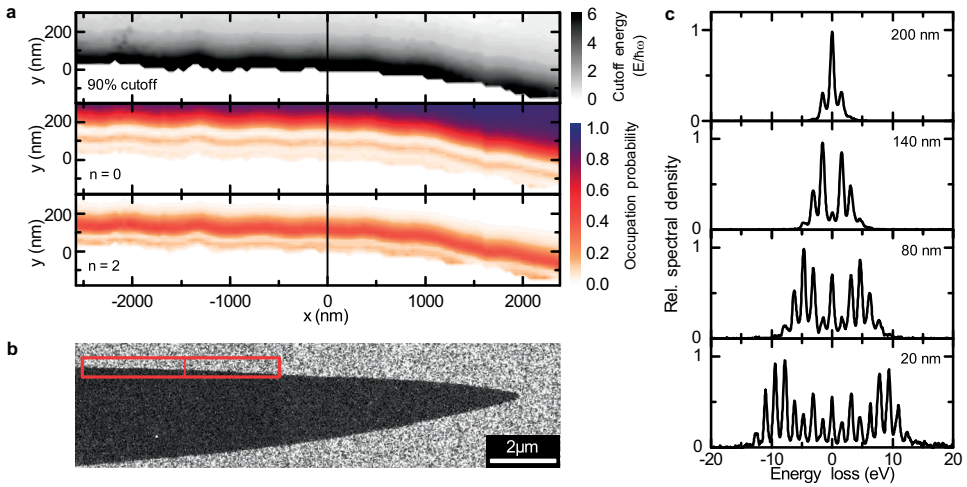
As shown in the main text, the sinusoidal phase modulation of the electron wavefunction by the interaction with the optical near-field leads to the formation of an attosecond pulse train after a certain distance of free propagation behind the interaction region. In the experiments, the electron pulse consists of a partially coherent ensemble of electrons, and we investigate here the robustness of the attosecond pulse train generation to an incoherent averaging over different coupling constants g and wavefunction evolutions with fluctuating initial energies. We find that an initial kinetic energy spread below the photon energy is fully sufficient for the formation of a clear attosecond pulse structure. Specifically, Extended Data Fig. 6 presents evolution maps of the electron pulse structure as a function of propagation distance, incoherently averaging simulations of pure states with an initial kinetic energy width of 0.1 eV each. In Extended Data Fig. 6a, b, the electron density is incoherently averaged over a range of kinetic energies $\Delta E = 0.7$ eV and 2.1 eV, respectively. At a range of 0.7 eV, the resulting electron density peak is practically indistinguishable from the ideal case of a pure state with 0.1 eV width (solid black line in Extended Data Fig. 6e). With increasing spread of the incoherent average, the peak begins to smear out, although its duration in the temporal focus is not notably enlarged even for a kinetic energy spread of 2.1 eV, three times larger than in the experiment.

An incoherent average over different coupling constants g experienced by the electrons within the electron beam area (lower row) has a different effect: for a small probing radius of 10 nm and a decay length of the coupling constant as used above, the peak width is not affected, but the depth of the temporal focus is broadened (Extended Data Fig. 6c). Increasing the probing radius to 50 nm (Extended Data Fig. 6d), that is, to a size substantially larger than in the experiment, the amplitude of the side lobes grows to ultimately affect the attosecond temporal resolution. In conclusion, the stability to perturbations in the coupling constant and the initial kinetic energy spread demonstrates that attosecond train generation will be observable under the given experimental conditions.

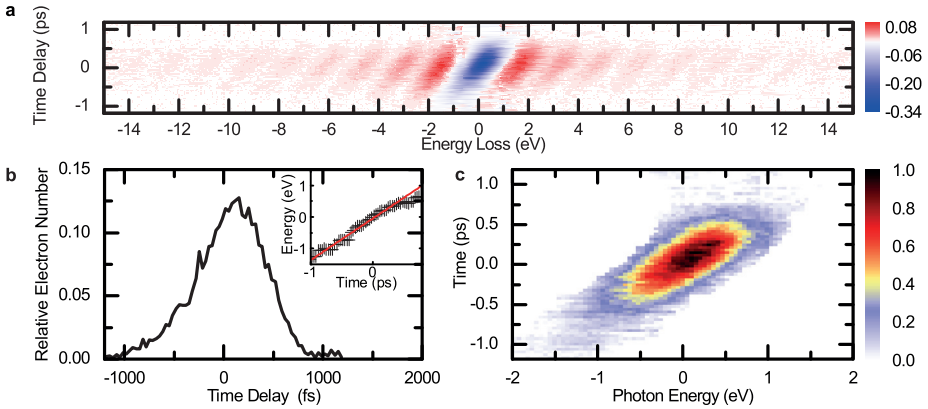
34. Park, S. T., Kwon, O.-H. & Zewail, A. H. Chirped imaging pulses in four-dimensional electron microscopy: femtosecond pulsed hole burning. *New J. Phys.* **14,** 053046 (2012).
35. Plemmons, D., Park, S. T., Zewail, A. H. & Flannigan, D. J. Characterization of fast photoelectron packets in weak and strong laser fields in ultrafast electron microscopy. *Ultramicroscopy* **146,** 97–102 (2014).
36. Egerton, R. F. Electron energy-loss spectroscopy in the TEM. *Rep. Prog. Phys.* **72,** 016502 (2009).
37. Schmidt, S. *et al.* Adiabatic nanofocusing on ultrasmooth single-crystalline gold tapers creates a 10-nm-sized light source with few-cycle time resolution. *ACS Nano* **6,** 6040–6048 (2012).
38. Ibe, J. *et al.* On the electrochemical etching of tips for scanning tunneling microscopy. *J. Vac. Sci. Technol. A* **8,** 3570–3575 (1990).
39. Asenjo-Garcia, A. & García de Abajo, F. J. Plasmon electron energy-gain spectroscopy. *New J. Phys.* **15,** 103021 (2013).
40. Park, S. T. & Zewail, A. H. Relativistic effects in photon-induced near field electron microscopy. *J. Phys. Chem. A* **116,** 11128–11133 (2012).
41. Schäfer, J., Lee, S.-C. & Kienle, A. Calculation of the near fields for the scattering of electromagnetic waves by multiple infinite cylinders at perpendicular incidence. *J. Quant. Spectrosc. Radiat. Transf.* **113,** 2113–2123 (2012).
42. Piazza, L. *et al.* Simultaneous observation of the quantization and the interference pattern of a plasmonic near-field. *Nature Comm.* **6,** 6407 (2015).



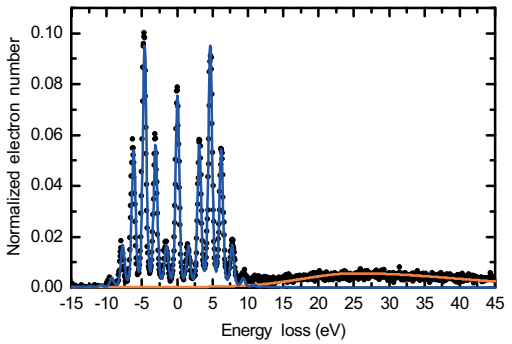
Extended Data Figure 1 | Experimental setup. Pulses from an amplified femtosecond (fs) laser system, at bottom left, are split into two optical beams. One of them is frequency-doubled in a β-barium borate (BBO) crystal and, after separation from the fundamental beam, focused (lens with numerical aperture 0.015, 50 cm focal length) onto the tungsten needle emitter (W tip) for the generation of electron probe pulses. The second beam (pump beam) is temporally stretched, attenuated and focused (lens with numerical aperture 0.014, 20 cm focal length) onto the sample within the TEM (angle of incidence, 55°). Relative timing between the electron probe and laser pump pulse is controlled by an optical delay stage. Optically-induced changes of the population of electron momentum states are recorded with an electron energy spectrometer. See Methods for details.



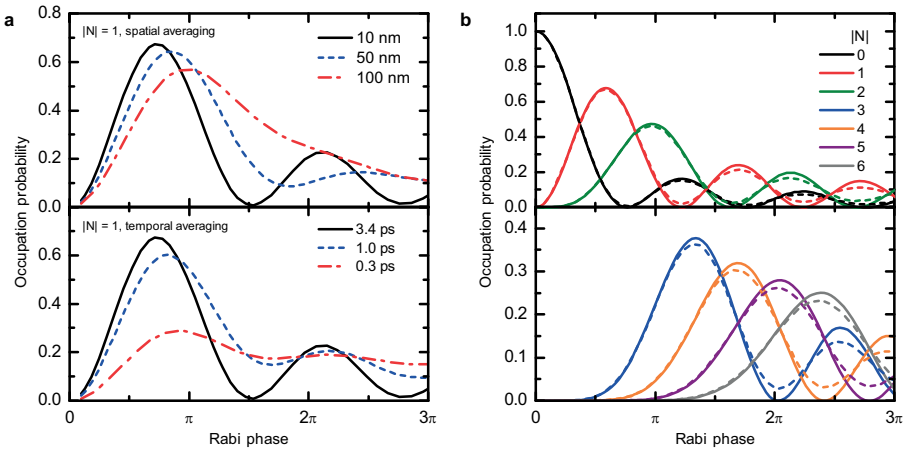
Extended Data Figure 2 | Spatial characterization of near-field scattering. **a**, Raster scan of the optically-induced electron energy gain and loss probability, characterized by the spectral cutoff (top panel) and the sideband populations of the zero loss peak (middle) and the second photon order (bottom). The field-dependent electron energy spectra shown in Fig. 2 of the main text were recorded at an x position indicated by the black line at the tip surface. A slow sample drift results in a scanning artefact in the y direction (jagged edge of the tip). For the results reported in the main text, a drift correction in the y direction was applied (see Methods section 'Data analysis and drift correction'). **b**, TEM image of gold tip. Red rectangle, scanning area displayed in **a**. **c**, Electron energy-loss spectra recorded along $x = 0$ with varying distance from the tip surface.



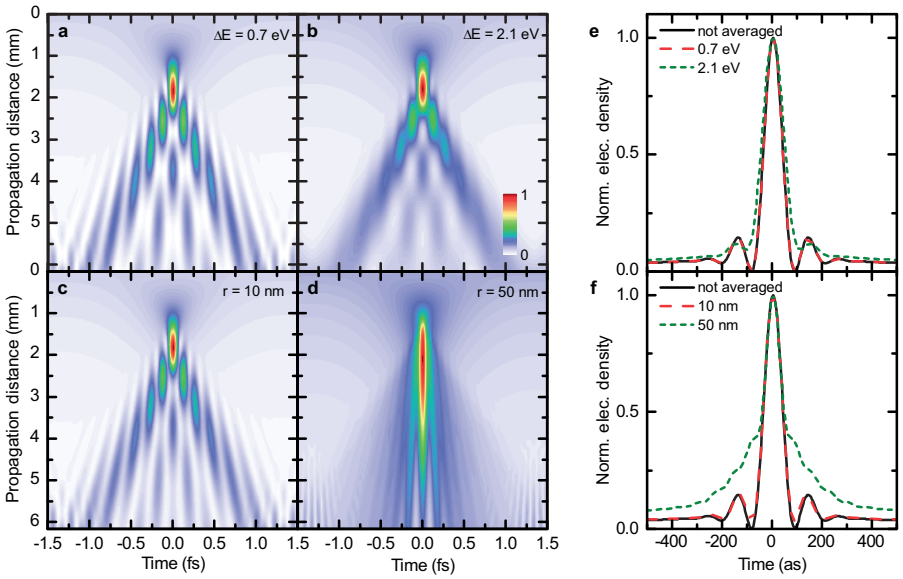
Extended Data Figure 3 | Pulse characterization by electron–photon cross-correlation. **a**, Differential electron energy-loss spectra as function of time delay (zero loss peak of width 1.3 eV subtracted; the colour scale shows the relative change of spectral density). **b**, Relative total scattering amplitude as function of time delay (inset, relative shift of photon sidebands with respect to zero loss peak). **c**, Energy- and time-resolved structure of the electron pulse (the colour scale shows the normalized electron density).



Extended Data Figure 4 | Evaluation of sideband populations. Example of electron energy spectrum (black dots) showing a number of photon sidebands and a weak low-loss plasmon contribution. Lines show fitted function used to extract sideband populations (blue) and the plasmon band (orange).



Extended Data Figure 5 | Influence of spatial and temporal averaging. **a**, Effect of electron beam size (top) and laser pulse duration (bottom) on the visibility of the Rabi oscillations in the order $|N| = 1$. For increasing electron beam size and decreasing laser pulse duration, the modulations are strongly damped. The black curves correspond to the experimental situation. **b**, Occupation probabilities of multiple spectral sidebands. Solid lines, *N*th-order Bessel functions. Dashed lines, numerical calculations accounting for temporal and spatial averaging in the experiments.



Extended Data Figure 6 | Robustness of attosecond pulse train generation. The influence of the electron beam's initial energy spread and lateral size on the temporal peak width of the generated pulse train are shown in the upper and lower rows, respectively. **a**–**d**, Evolution of the electron density as a function of propagation distance after the interaction with the optical near-field, incoherently averaged over the initial kinetic energy distribution (**a**, **b**) or the finite probing-area of the electron beam (**c**, **d**). A corresponding line profile at the propagation distance where the electron density peaks are at their maximum and form an attosecond pulse train is shown in **e** and **f**. For the experimental parameters used in this work (energy spread $\Delta E = 0.7$ eV FWHM and electron beam radius $r = 10$ nm), the peak width remains nearly unchanged as compared to the ideal (not averaged) case.

LETTER

doi:10.1038/nature14400

Global carbon export from the terrestrial biosphere controlled by erosion

Valier Galy[1], Bernhard Peucker-Ehrenbrink[1] & Timothy Eglinton[1,2]

Riverine export of particulate organic carbon (POC) to the ocean affects the atmospheric carbon inventory over a broad range of timescales[1–5]. On geological timescales, the balance between sequestration of POC from the terrestrial biosphere and oxidation of rock-derived (petrogenic) organic carbon sets the magnitude of the atmospheric carbon and oxygen reservoirs[6,7]. Over shorter timescales, variations in the rate of exchange between carbon reservoirs, such as soils and marine sediments, also modulate atmospheric carbon dioxide levels[1]. The respective fluxes of biospheric and petrogenic organic carbon are poorly constrained, however, and mechanisms controlling POC export have remained elusive, limiting our ability to predict POC fluxes quantitatively as a result of climatic or tectonic changes. Here we estimate biospheric and petrogenic POC fluxes for a suite of river systems representative of the natural variability in catchment properties. We show that export yields of both biospheric and petrogenic POC are positively related to the yield of suspended sediment, revealing that POC export is mostly controlled by physical erosion. Using a global compilation of gauged suspended sediment flux, we derive separate estimates of global biospheric and petrogenic POC fluxes of 157^{+74}_{-50} and 43^{+61}_{-25} megatonnes of carbon per year, respectively. We find that biospheric POC export is primarily controlled by the capacity of rivers to mobilize and transport POC, and is largely insensitive to the magnitude of terrestrial primary production. Globally, physical erosion rates affect the rate of biospheric POC burial in marine sediments more strongly than carbon sequestration through silicate weathering. We conclude that burial of biospheric POC in marine sediments becomes the dominant long-term atmospheric carbon dioxide sink under enhanced physical erosion.

The atmosphere is a small reservoir of carbon in comparison with rocks, soils, the biosphere and the ocean[1]. Its size is therefore sensitive to small imbalances in the exchange of C with and between these larger reservoirs. Over long timescales, the continental biosphere is mostly at equilibrium with the atmosphere, because most C fixed by terrestrial photosynthesis is quickly returned to the atmospheric reservoir through respiration[1]. However, rivers deliver to the oceans a fraction of this net primary production (NPP) as POC and dissolved organic carbon (DOC)[2–5]. Although most DOC is quickly returned to the atmosphere through oxidation in estuaries and the ocean, a significant fraction of riverine POC is buried in marine sediments and stored over long timescales. This 'leakage' of carbon from the biosphere–atmosphere loop represents a net sequestration of atmospheric C (ref. 6). Rivers also transfer POC from the rock reservoir (petrogenic organic carbon, OC) to marine sediments, thereby transferring C between two reservoirs disconnected from the atmosphere[7]. During this transfer, oxidation of petrogenic OC represents another leakage of C, in this case towards the atmosphere[8]. The nature and efficiency of riverine export of POC to the ocean thus fundamentally affect the long-term atmospheric C inventory. Despite its importance, global riverine export of POC to the ocean has until now remained poorly constrained[2–5]. In particular, the respective global fluxes of biospheric and petrogenic POC remain largely unconstrained. More importantly, the sensitivities and relative magnitudes of global biospheric and petrogenic POC export are not well defined, impeding our ability to quantitatively predict POC fluxes and their impact on the long-term global C cycle under different forcing scenarios.

Developing accurate constraints on fluvial transfer of biospheric POC requires the quantification of, and correction for, petrogenic OC in river sediment. Although the presence of petrogenic OC in river sediments and river-dominated margin sediments has been inferred for decades, its direct and unambiguous detection is quite recent[7,9,10]. Consequently, few quantitative reconstructions of petrogenic OC fluxes exist. However, those reconstructions encompass such diverse river systems as the Amazon, Taiwanese rivers and the Ganges–Brahmaputra system[7,8,11]. Radiocarbon (^{14}C) measurements have provided key constraints on petrogenic OC concentrations and fluxes. Exploiting the absence of ^{14}C in petrogenic OC and its presence in biospheric OC, ^{14}C measurements on riverine POC combined with additional constraints on the chemical composition of either constituent allow these two key constituents to be differentiated[7,12–18]. Here we use published POC compositional data (including ^{14}C measurements) to derive a direct, global estimate of the petrogenic OC flux (Methods). These new results, together with published petrogenic OC fluxes, provide a unique compilation of riverine export of petrogenic and biospheric POC to the ocean from 43 river systems that account for 20% of the sediment discharge to the oceans.

Petrogenic OC is an integral component of sedimentary and other rocks. Its export by river systems is therefore tightly linked to that of sediments. Indeed, both Komada *et al.*[17] and Hilton *et al.*[16] have shown that yields of petrogenic POC (the petrogenic POC flux normalized to catchment area) are positively correlated with corresponding yields of suspended sediment in the Santa Clara River (California) and Taiwanese rivers, respectively, with the latter property serving as a measure of spatially averaged physical erosion rate. Our data set extends this observation to a broad range of river systems, covering more than four orders of magnitude in both catchment size and sediment yield (Supplementary Table 1). These data broadly follow a power-law relationship characterized by a power exponent close to unity (1.11 ± 0.13) (errors are 1 s.d.; see Methods). Petrogenic POC yield thus varies roughly linearly with sediment yield, implying that the behaviour of petrogenic OC during erosion and transport is similar to that of the other mineral phases (Fig. 1). In particular, this finding implies a generally uniform depth-distribution of petrogenic OC in soils and rocks. The average petrogenic POC concentration in river sediments, however, varies considerably (0.02% to 0.6%). This can be explained by variations in the average petrogenic OC content of rocks and/or by variable oxidation of petrogenic OC during sediment transfer to the ocean. Recent studies of large river basins with extensive floodplains (Ganges–Brahmaputra and Amazon) indicate that up to 50% of petrogenic POC initially present in rocks can be oxidized during sediment transport and temporary storage in intermediate reser-

[1]Woods Hole Oceanographic Institution, Department of Marine Chemistry and Geochemistry, 360 Woods Hole Road, Woods Hole, Massachusetts 02543, USA. [2]Geological Institute, Department of Earth Sciences, Sonneggstrasse 5, Eidgenössische Technische Hochschule, 8092 Zürich, Switzerland.

voirs[7,8]. In contrast, petrogenic OC is very efficiently preserved in fluvial systems characterized by rapid sediment transfer to the ocean, such as Taiwanese rivers[19] and the Eel River (California)[20]. Enhanced oxidation of petrogenic OC could explain up to about 50% of the observed order-of-magnitude difference in petrogenic OC concentration. In addition, average petrogenic OC concentrations in rocks have been reported to vary from catchment to catchment by at least an order of magnitude[10,21]. Together, these observations suggest that whereas initial contents in rocks and subsequent oxidation of petrogenic OC during sediment transport together are the dominant controls of petrogenic OC concentration in river sediments, sediment yield—that is, physical erosion rate—is the primary control on petrogenic OC export efficiency.

Unlike petrogenic OC, biospheric POC is not an indigenous mineral component of the sediment; instead, it is added during vegetation growth, soil formation and processes associated with the movement of materials from source to sink (for example landsliding or overland flow). The controls on its behaviour could therefore differ from those affecting the mineral load. Because it is the other component of riverine POC, biospheric OC fluxes can be obtained by subtracting petrogenic OC contributions from riverine POC fluxes. Here we use a compilation of riverine POC fluxes from 70 river systems (Supplementary Tables 1 and 2), covering 42.7×10^6 km^2 (that is, 40% of the total exorheic continental area) and accounting for about 45% of the global freshwater discharge to the ocean. For 27 of these river systems we lack direct estimates of petrogenic OC contributions. We therefore use the relationship between sediment yield and petrogenic OC yield (Fig. 1) to estimate petrogenic contributions to POC in these systems (Methods). This permits the calculation of biospheric POC fluxes and yields for the entire set of 70 river systems. Our data show that biospheric POC yield is positively correlated with suspended sediment yield, following a unique power-law relationship ($r^2 = 0.78$) (Fig. 2). Although relationships between riverine POC and suspended sediment concentrations have been reported previously[2], we extend this observation to biospheric POC, accounting for its dilution by petrogenic OC at high sediment yield. The singularity of the relationship between biospheric POC and sediment yield is remarkable, considering the very broad range of climate, vegetation, geomorphology and anthropogenic disturbance characterizing the drainage basins considered. It implies that, globally, the rate of biospheric POC export is primarily controlled by sediment export processes. The exponent of the power relationship between biospheric POC and sediment yield is significantly smaller than 1 (0.56 ± 0.03), reflecting an increasing dilution of biospheric POC by mineral phases (that is, decreasing biospheric POC concentrations) at high sediment yield. This reflects the well-documented decrease in OC concentration with depth in soil profiles, which results in an increase in POC stock with depth that globally follows a power law characterized by an exponent of 0.4 (ref. 22). In general, at low sediment yield, erosion proceeds mostly by means of overland flow, exporting surface material (such as soil litter and O horizons) characterized by high biospheric POC concentrations. Conversely, at high sediment yield, erosion proceeds by means of deep gully erosion and/or landslides, thereby lowering the overall biospheric POC concentration of the eroded material[23].

To further evaluate to what extent productivity and associated soil OC content control biospheric POC export, we used the MOD17 database[24] to extract basin-scale estimates of NPP for 40 systems. Biospheric OC yield and NPP are weakly positively correlated (power law with exponent of 1.16 and $r^2 = 0.30$; Extended Data Fig. 1), suggesting that productivity does not impose a strong control on biospheric OC yield. However, the calculated fractions of the NPP exported from catchments by rivers vary by more than three orders of magnitude (0.01% to 2.1%) and are positively correlated with the suspended sediment yield, following a power-law relationship (Fig. 3) that is characterized by an exponent (0.50 ± 0.05) statistically identical to that of the relationship defined by biospheric POC yield. We therefore conclude that the rate of biospheric POC export depends primarily on the capacity of rivers to mobilize and transport POC out of catchments, rather than on POC production within the watershed. Our data set also reveals significant secondary variations (Figs 2 and 3), suggesting the existence of additional control mechanisms. Among the possible mechanisms, we postulate that, for a given sediment yield, higher frequency and/or deeper landslides result in decreased biospheric POC yield, illustrating the critical role of physical erosion processes in biospheric POC export. Other possible mechanisms include the sorption of DOC onto mineral phases in sediment-starved rivers (for example the Congo river) that are often characterized by high DOC concentrations. However, the lack of correlation between DOC concentration and biospheric POC yield suggests that sorption of DOC does not exert a significant control on the efficiency of biospheric POC export. The increasing relative variance of biospheric POC yield at a low concentration of suspended sediment—conditions that promote aquatic primary production—suggests that within-river biological productivity could explain part of the secondary variability of biospheric POC yield. Transient non-steady-state erosion of soil and biosphere reservoirs can also introduce some variability in biospheric POC yield that might not be accounted for by either sediment yield or NPP. Finally, human activities such as damming and agriculture currently affect virtually all modern river systems and probably influence biospheric POC yields.

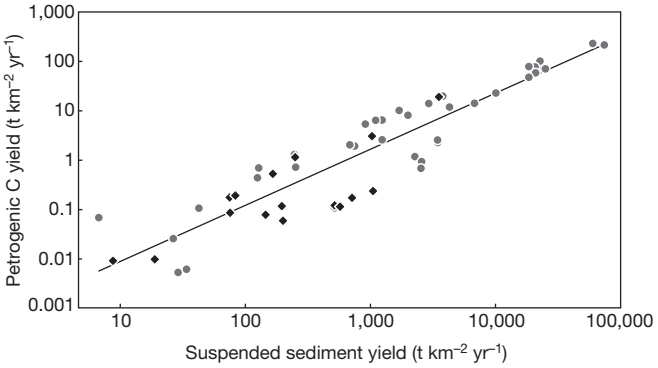


Figure 1 | Relationship between petrogenic OC yield (Y_{petro}) and suspended sediment yield (Y_{sed}). Catchments larger (black diamonds) or smaller (grey dots) than 100,000 km^2 plot on the same trend. Most of the variability in petrogenic OC concentration derives from variable initial OC concentrations in rocks and petrogenic OC oxidation during sediment transport. The regression line is $Y_{petro} = 0.0007 Y_{sed}^{1.11}$; $r^2 = 0.82$; $P < 0.001$.

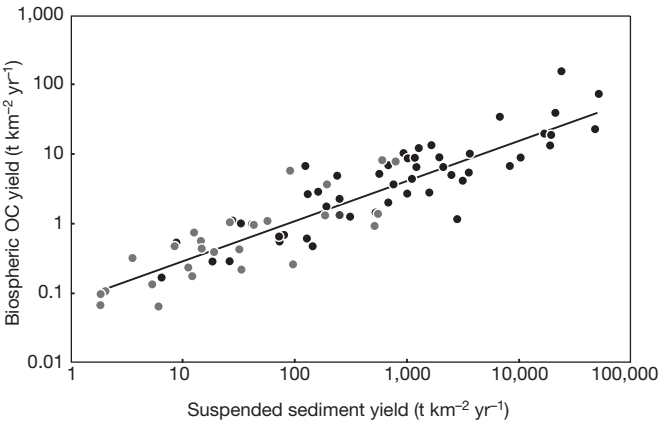


Figure 2 | Relationship between biospheric POC yield (Y_{bios}) and suspended sediment yield. Data obtained by subtracting measured petrogenic OC fluxes from riverine POC fluxes (black dots) and those obtained using petrogenic OC fluxes inferred from the relationship shown in Fig. 1 (grey dots) plot on the same trend. The regression line is $Y_{bios} = 0.081 Y_{sed}^{0.56}$; $r^2 = 0.78$; $P < 0.001$.

The strong relationships between suspended sediment yield and POC yield (both petrogenic and biospheric) allow global POC fluxes to be inferred from the better-constrained global sediment flux. Here we use recent estimates of the suspended sediment flux to the ocean[25–27] to estimate global biospheric and petrogenic POC fluxes. We use a global suspended sediment flux of 19,000 ± 500 Mt yr^{-1} and a corresponding suspended sediment yield of 176 ± 8 t km^{-2} yr^{-1} (Methods). In turn, using the overall relationship between suspended sediment yield and biospheric POC yield (Fig. 2), we estimate a global biospheric POC yield of $1.46^{+0.68}_{-0.47}$ t C km^{-2} yr^{-1}, which translates into a global biospheric POC flux of 157^{+74}_{-50} Mt C yr^{-1}. A similar approach applied to the relationship between suspended sediment yield and the fraction of the NPP exported by rivers (Fig. 3) gives a global value of 0.18% of terrestrial NPP being exported to the ocean. Combined with a mean estimate of the global terrestrial NPP of 77.6 Gt C yr^{-1} (ref. 24), this gives a global biospheric POC flux of 140^{+96}_{-57} Mt C yr^{-1} that is statistically identical to our estimate based on biospheric POC yield data. Annually, about 0.02% of the total mass of C present in the atmosphere is thus transferred to the ocean as POC, showing that biospheric POC sequestration can affect the size of the atmospheric reservoir over timescales as short as 10^3–10^4 years.

The dependence of petrogenic OC yield on OC concentrations in rocks complicates the estimation of the global petrogenic POC flux from our estimate of global suspended sediment yield. Ideally, the distribution of rocks characterized by variable OC concentrations as well as intrinsic geomorphic characteristics (such as the size of the floodplain and the spatial distribution of physical erosion) need to be taken into account, because they both exert strong control on petrogenic OC yields. In the absence of such a model, we can only assume that the 43 river systems that we characterized are representative of the natural variability (that is, rock types and catchment morphology). Using our global suspended sediment yield of 176 ± 8 t km^{-2} yr^{-1}, we estimate a global petrogenic POC flux of 43^{+61}_{-25} Mt C yr^{-1} (Methods).

Finally, we derive a combined global flux of terrestrial POC to the ocean of 200^{+135}_{-75} Mt C yr^{-1}, of which about 80% and 20% are biospheric and petrogenic POC, respectively. This direct estimate of these two fluxes provides an assessment of the magnitude of POC transfer from the terrestrial biosphere to the ocean, and reveals the global significance of petrogenic OC as a component of POC export by rivers to the ocean. However, these fluxes do not take bedload transport into account and must therefore represent a lower bound of actual petrogenic and biospheric POC fluxes to the ocean. Indeed, bedload material can be dominated either by petrogenic OC[7,8] or biospheric POC[28], implying that bedload transport contributes globally to the fluvial export of both petrogenic and biospheric POC.

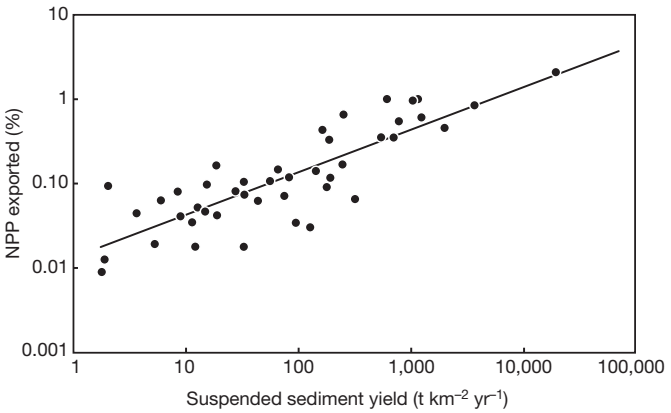


Figure 3 | Relationship between the proportion of NPP exported annually (NPP_{exp}) and suspended sediment yield. Normalization of biospheric POC export to NPP does not remove its dependence on suspended sediment yield, illustrating the overarching control exerted by physical erosion on biospheric POC export. The regression line is $NPP_{exp} = 0.013 Y_{sed}^{0.50}$; $r^2 = 0.71$; $P < 0.001$.

On geological timescales, petrogenic and biospheric OC have opposing roles in the global C cycle: the net transfer of C between atmospheric and terrestrial reservoirs is set by the balance between petrogenic OC oxidation and biospheric POC burial. We show that the rate of both petrogenic and biospheric POC export from the continents is controlled primarily by the rate of sediment export; that is, physical erosion. The preservation of both petrogenic and biospheric POC in the ocean is up to three times higher at high physical erosion rates[10,16,19]. Thus, increased physical erosion rates favour efficient transfer and burial of biospheric POC coupled with enhanced preservation (that is, decreased oxidation) of petrogenic POC, both acting to limit the return of carbon to the atmospheric reservoir. The small fraction of NPP exported even at very high erosion rates (a few per cent; Fig. 3) suggests that enhanced export of biospheric POC is sustainable over long timescales, as the terrestrial biospheric OC reservoir is continuously replenished by photosynthetic C fixation. Using available estimates of terrestrial POC burial efficiency, we show that biospheric POC burial yield is positively correlated with sediment yield (Extended Data Fig. 2). Globally, the rate of C sequestration through silicate weathering has a weaker sensitivity to sediment yield (Extended Data Fig. 2) as a result of kinetic limitation of weathering reactions at high physical erosion rates[29]. Biospheric POC burial is thus predicted to become the dominant long-term atmospheric CO_2 sink under a fourfold increase in global physical erosion rate at constant temperature. We conclude that tectonic and climatic forcing of physical erosion favours biospheric POC sequestration over silicate weathering.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September 2014; accepted 9 March 2015.

1. Sarmiento, J. & Gruber, N. in *Ocean Biogeochemical Dynamics* (eds Sarmiento, J. & Gruber, N.) 392–453 (Princeton Univ. Press, 2006).
2. Ludwig, W., Probst, J.-L. & Kempe, S. Predicting the oceanic input of organic carbon by continental erosion. *Glob. Biogeochem. Cycles* **10,** 23–41 (1996).
3. Schlünz, B. & Schneider, R. R. Transport of terrestrial organic carbon to the oceans by rivers: re-estimating flux and burial rates. *Int. J. Earth Sci.* **88,** 599–606 (2000).
4. Meybeck, M. in *Interactions of C, N, P and S: Biogeochemical Cycles and Global Change* (eds Wollast, R., Mackenzie, F. T. & Chou, L.) 163–193 (Springer, 1993).
5. Degens, E. T., Kempe, S. & Richey, J. E. *Biogeochemistry of Major World Rivers* (Wiley, 1991).
6. Berner, R. A. Burial of organic carbon and pyrite sulfur in the modern ocean: its geochemical and environmental significance. *Am. J. Sci.* **282,** 451–473 (1982).
7. Galy, V., Beyssac, O., France-Lanord, C. & Eglinton, T. I. Recycling of graphite during Himalayan erosion: a geological stabilization of carbon in the crust. *Science* **322,** 943–945 (2008).
8. Bouchez, J. *et al.* Oxidation of petrogenic organic carbon in the Amazon floodplain as a source of atmospheric CO_2. *Geology* **38,** 255–258 (2010).
9. Blair, N. E., Leithold, E. L. & Aller, R. C. From bedrock to burial: the evolution of particulate organic carbon across coupled watershed–continental margin systems. *Mar. Chem.* **92,** 141–156 (2004).
10. Galy, V. *et al.* Efficient organic carbon burial in the Bengal fan sustained by the Himalayan erosional system. *Nature* **450,** 407–410 (2007).
11. Hilton, R. G. *et al.* Climatic and geomorphic controls on the erosion of terrestrial biomass from subtropical mountain forest. *Glob. Biogeochem. Cycles* **26,** GB3014 (2012).
12. Bouchez, J. *et al.* Source, transport and fluxes of Amazon River particulate organic carbon: insights from river sediment depth-profiles. *Geochim. Cosmochim. Acta* **133,** 280–298 (2014).
13. Drenzek, N. *et al.* A new look at old carbon in active margin sediments. *Geology* **37,** 239–242 (2009).
14. Galy, V. & Eglinton, T. I. Protracted storage of biospheric carbon in the Ganges–Brahmaputra basin. *Nature Geosci.* **4,** 843–847 (2011).
15. Hilton, R. G. *et al.* Tropical-cyclone-driven erosion of the terrestrial biosphere from mountains. *Nature Geosci.* **1,** 759–762 (2008).
16. Hilton, R. G., Galy, A., Hovius, N. & Horng, M. J. Efficient transport of fossil organic carbon to the ocean by steep mountain rivers: an orogenic carbon sequestration mechanism. *Geology* **39,** 71–74 (2011).
17. Komada, T., Druffel, E. R. M. & Trumbore, S. E. Ocanic export of relict organic carbon by small mountainous rivers. *Geophys. Res. Lett.* **31,** 1–4 (2004).

18. Leithold, E. L., Blair, N. E. & Perkey, D. W. Geomorphologic controls on the age of particulate organic carbon from small mountainous and upland rivers. *Glob. Biogeochem. Cycles* **20,** GB3022 (2006).
19. Kao, S.-J. *et al.* Preservation of terrestrial organic carbon in marine sediments offshore Taiwan: mountain building and atmospheric carbon dioxide sequestration. *Earth Surf. Dyn.* **2,** 127–139 (2014).
20. Blair, N. E. *et al.* The persistence of memory: the fate of ancient sedimentary organic carbon in a modern sedimentary system. *Geochim. Cosmochim. Acta* **67,** 63–73 (2003).
21. Hilton, R. G., Galy, A., Hovius, N., Horng, M. J. & Chen, H. E. The isotopic composition of particulate organic carbon in mountain rivers of Taiwan. *Geochim. Cosmochim. Acta* **74,** 3164–3181 (2010).
22. Jobbagy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10,** 423–436 (2000).
23. Hilton, R. G., Meunier, P., Hovius, N., Bellingham, P. J. & Galy, A. Landslide impact on organic carbon cycling in a temperate montane forest. *Earth Surf. Process. Landf.* **36,** 1670–1679 (2011).
24. Zhao, M., Nemani, Z. & Running, S. (ed. NASA).
25. Milliman, J. D. & Farnsworth, K. *River Discharge to the Coastal Ocean: a Global Synthesis* (Cambridge Univ. Press, 2011).
26. Peucker-Ehrenbrink, B. Land2Sea database of river drainage basin sizes, annual water discharges, and suspended sediment fluxes. *Geochem. Geophys. Geosyst.* **10,** Q06014 (2009).
27. Larsen, I. J., Montgomery, D. R. & Greenberg, H. M. The contribution of mountains to global denudation. *Geology* **42,** 527–530 (2014).
28. Bianchi, T. S., Galler, J. J. & Allison, M. A. Hydrodynamic sorting and transport of terrestrially derived organic carbon in sediments of the Mississippi and Atchafalaya Rivers. *Estuar. Coast. Shelf Sci.* **73,** 211–222 (2007).
29. West, A. J., Galy, A. & Bickle, M. Tectonic and climatic controls on silicate weathering. *Earth Planet. Sci. Lett.* **235,** 211–228 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Godderis, J. Hemingway and G. Soulet for comments on early versions of the manuscript. G. Fiske generated the NPP data. Support for this project was provided by US National Science Foundation (NSF) grant OCE-0851015 (to B.P.-E., T.E. and V.G.), NSF grant OCE-0928582 (to V.G. and T.E.) and Swiss National Science Foundation grant 200021_140850 (to T.E.).

Author Contributions V.G. designed the study, performed the analysis and drafted the manuscript with inputs from B.P.-E. and T.E.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.G. (vgaly@whoi.edu).

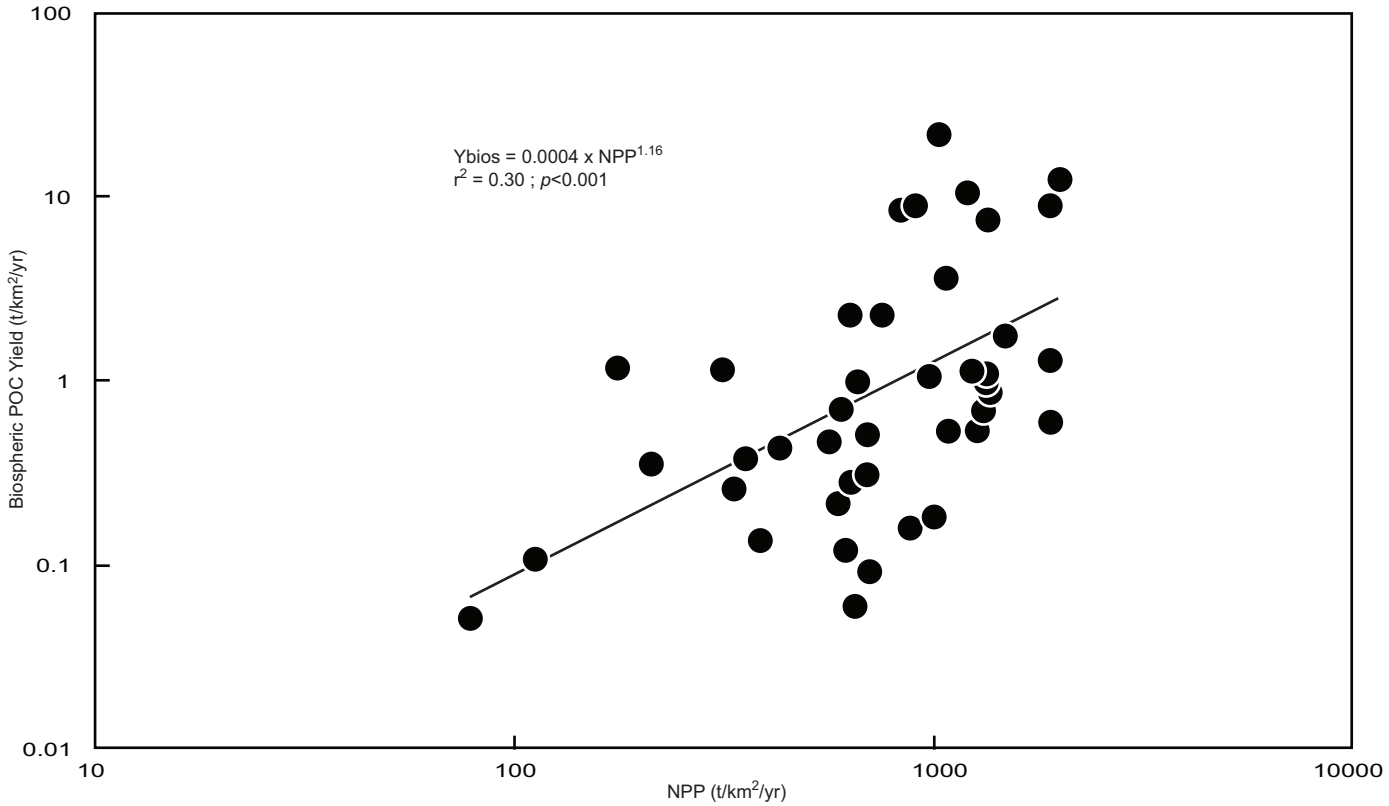
METHODS

Quantitative apportionment of biospheric and petrogenic POC in river sediments. Several methods have recently been used to apportion petrogenic and biospheric POC quantitatively in river sediments. All of these methods are based on the unique property of petrogenic OC that it does not contain any ^{14}C, whereas biospheric POC does. However, the ^{14}C content of biospheric POC is difficult to predict a priori, because it depends on a complex array of processes such as physical erosion, soil formation and the dynamics of the biosphere itself. Therefore, in most cases, additional constraints on the chemical composition of biospheric and/or petrogenic OC are needed. Galy *et al.*[12] proposed that the different hydrodynamic properties of petrogenic and biospheric POC in large streams and rivers enables the use of bulk radiocarbon measurements on POC in suspended and bed sediments collected along depth profiles to estimate both petrogenic OC concentration and biospheric POC ^{14}C content. The underlying assumption is that petrogenic OC is uniformly distributed in the water column (owing to its physical properties and size distribution), whereas the relative concentration of biospheric POC—which is preferentially associated with fine-grained sediments—decreases predictably with depth. This approach has been tested and yields robust results for the Ganges–Brahmaputra[7,14], Amazon[8,12] and Fraser[30] rivers. Here we reanalyse published ^{14}C data and show that the approach developed by Galy *et al.*[7] for suspended sediment depth profiles can also be applied to size-fractionated and density-fractionated fluvial sediments as well as to suspended sediments sampled across a wide range of flow conditions. The ^{14}C content of density-fractionated sediments derived from the Mississippi river[31] and of time-series suspended sediments of the Ishikari river[32] provide excellent examples (Extended Data Figs 3 and 4). This type of binary mixing approach is, however, not always appropriate, either because adequate data may not be available or because of non-systematic behaviour of biospheric POC in the water column. Indeed, biospheric POC often reflects the mixing of several components such as soil and fresh plant debris, which may associate with the mineral load in different ways and have different hydrodynamic properties. Here the Mackenzie river provides a good example. Recently fixed plant-derived biospheric OC and very old permafrost-derived biospheric OC are preferentially associated with coarse and fine sediments, respectively, undermining the binary mixing approach[33]. In these circumstances, additional constraints on the composition of the petrogenic and biospheric end-members are needed. These can be obtained by using bulk (for example N/C, $\delta^{13}C$ or $\delta^{15}N$) or compound-specific (^{14}C) data, as demonstrated for Taiwanese rivers[21], the Eel River[13] and the Mackenzie River[34]. Once a priori compositions of the petrogenic and different biospheric end-members have been established, simple mixing models can be used to apportion petrogenic and biospheric POC quantitatively. We used a combination of the above-mentioned techniques to evaluate petrogenic OC concentrations in river sediments by using published and newly acquired POC characterizations. Supplementary Table 1 summarizes the sources of the data, the methods employed and the results obtained.

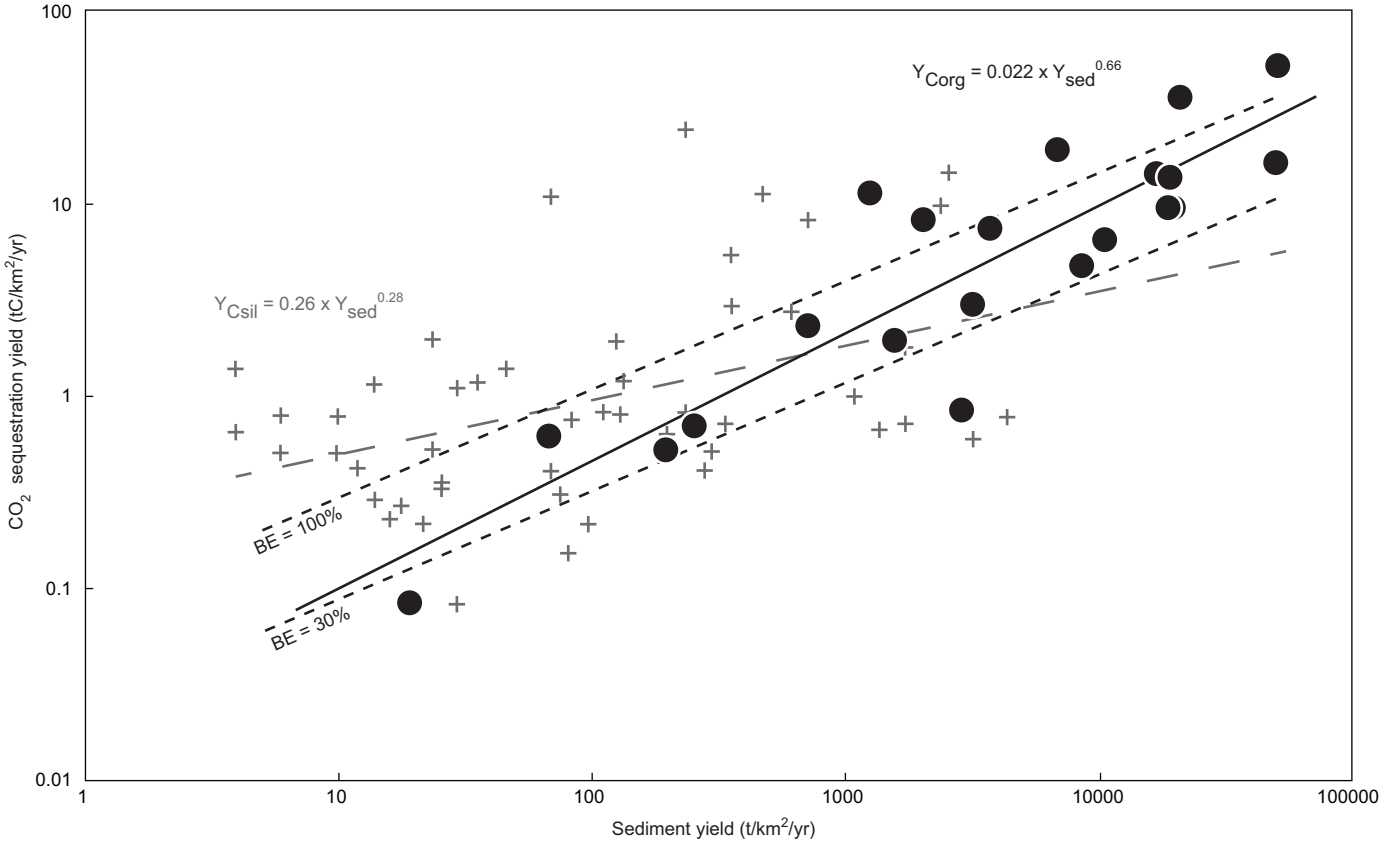
In some cases, POC fluxes have been measured but adequate data to quantify petrogenic and biospheric POC are not available. In these cases (27 rivers; Supplementary Table 2) we estimate the petrogenic OC yield by using the relationship between suspended sediment yield and petrogenic OC yield defined by all rivers for which we could quantify petrogenic and biospheric OC (43 rivers; Fig. 1 and Supplementary Table 1). The significant scatter around this relationship introduces uncertainty in the estimation of petrogenic OC yield from suspended sediment yield. We therefore estimated the uncertainty of calculated petrogenic OC yield on the basis of the uncertainty of the relationship between suspended sediment yield and petrogenic OC yield. Biospheric POC yield is obtained by subtracting inferred petrogenic OC yield from riverine POC yield. The relationship between biospheric POC yield and suspended sediment yield is identical when the two groups of data—that is, data determined from geochemical characterization versus data inferred from the relationship between suspended sediment yield and petrogenic OC yield—are considered separately (Fig. 2). Specifically, the exponents and multiplying terms of the two relationships are statistically identical (within ± 1 s.d.): 0.51 ± 0.04 and 0.59 ± 0.09 for the exponents, and -0.93 ± 0.13 and -1.19 ± 0.15 for the multiplying term. This shows that inferring petrogenic OC yield from the relationship between suspended sediment yield and petrogenic OC yield does not introduce a systematic bias in the determination of biospheric POC yield.

Estimating global riverine petrogenic and biospheric POC fluxes. Robust relationships between both petrogenic and biospheric POC yields and suspended sediment yield enable the use of global suspended sediment yield to estimate global petrogenic and biospheric POC fluxes. Global suspended sediment fluxes have been the subject of extensive research over the past several decades. Suspended sediment fluxes for individual gauged rivers have been compiled[25,26,35,36] and used to derive estimates of global suspended sediment fluxes. Usually, fluxes were first extrapolated regionally (for example by grouping rivers according to the oceanic basin they drain into) to account for regional differences in average suspended sediment yield (for example very high in small ocean islands versus small in the Russian Arctic). Global fluxes were then obtained by summing the regional fluxes for all areas draining into oceanic basins (in other words, endorheic systems are excluded from the global estimate). Peucker-Erhenbrink[26] and Milliman and Farnsworth[25] have provided the two most comprehensive recent compilations of gauged suspended sediment fluxes. They estimated the global suspended sediment flux to the ocean at 18,548 and 19,000 Mt yr^{-1}, respectively. Recently, Larsen *et al.*[27] estimated the global exorheic denudation flux at 19,000 Mt yr^{-1}, using an empirical denudation model based on the strong relationship between denudation rates and topography. Here we use a global suspended sediment flux of $19{,}000 \pm 500$ Mt yr^{-1}. Graham *et al.*[37] estimated the global exorheic land surface at 110×10^6 km^2, whereas Syvitski *et al.*[38] proposed a slightly lower estimate of 106×10^6 km^2. The difference between these two estimates probably derives from differences in corrections for endorheic drainage areas[26]. Here we use an average value of $(108 \pm 2) \times 10^6$ km^2, resulting in an estimated global suspended sediment yield of 176 ± 8 t km^{-2} yr^{-1}. It should be noted that this value attempts to correct, as far as possible, for recent damming of river basins, because pre-dam sediment fluxes were used to estimate the global suspended sediment fluxes whenever possible both by Peucker-Ehrenbrink *et al.*[26] and by Milliman and Farnsworth[25]. Finally, we use the calculated value for global suspended sediment yield (176 ± 8 t km^{-2} yr^{-1}) and our relationships between petrogenic and biospheric POC yields and suspended sediment yield to derive estimates of the global petrogenic and biospheric POC fluxes. To estimate the uncertainties associated with calculated global fluxes we first use Monte Carlo simulations (10,000 repetitions) to account for the uncertainty associated with the determination of the relationships between petrogenic and biospheric POC yields and suspended sediment yield. Then we propagate the uncertainty (± 8 t km^{-2} yr^{-1}) in the global suspended sediment yield. Last, we compare the lower bound of the calculated global petrogenic and biospheric POC fluxes with the sum of all petrogenic and biospheric POC fluxes, respectively, and use the highest of the two values as the lower bound of our final estimate.

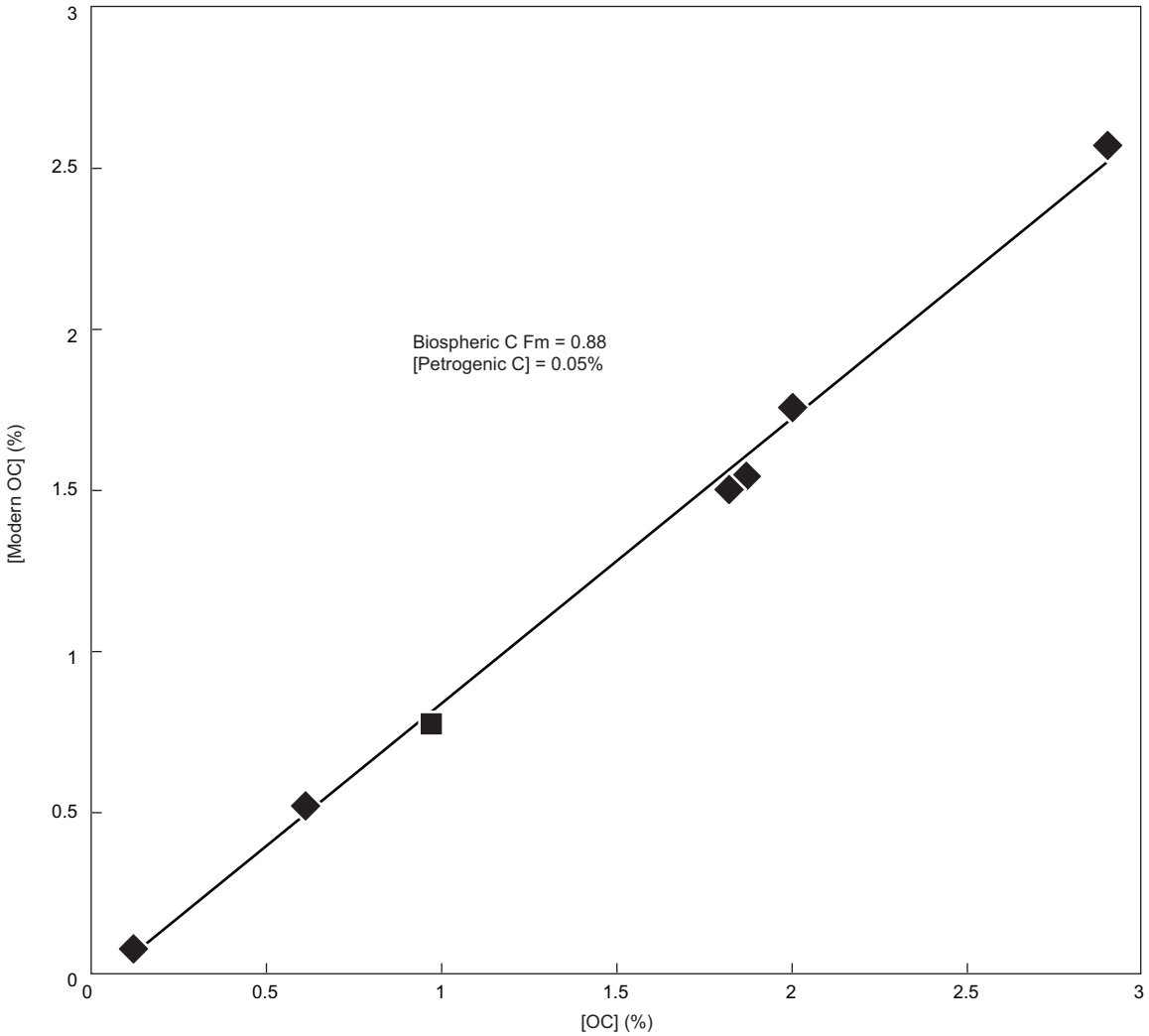
30. Voss, B. M. *Spatial and Temporal Dynamics of Biogeochemical Processes in the Fraser River, Canada: a Coupled Organic–Inorganic Perspective.* PhD thesis, Massachussetts Institute of Technology and Woods Hole Oceanographic Institution (2014).
31. Wakeham, S. G. *et al.* Partitioning of organic matter in continental margin sediments among density fractions. *Mar. Chem.* **115,** 211–225 (2009).
32. Alam, M. J., Nagao, S., Aramaki, T., Shibata, Y. & Yoneda, M. Transport of particulate organic matter in the Ishikari River, Japan during spring and summer. *Nuclear Instrum. Meth. Phys. Res. B* **259,** 513–517 (2007).
33. Hilton, R. G. *et al.* Erosion of organic carbon in the Arctic as a geological carbon dioxide sink. *Nature* (submitted).
34. Drenzek, N. J., Montlucon, D. B., Yunker, M. B., Macdonald, R. W. & Eglinton, T. I. Constraints on the origin of sedimentary organic carbon in the Beaufort Sea from coupled molecular ^{13}C and ^{14}C measurements. *Mar. Chem.* **103,** 146–162 (2007).
35. Milliman, J. D. & Meade, R. H. World delivery of river sediment to the oceans. *J. Geol.* **1,** 1–21 (1983).
36. Milliman, J. D. & Syvitski, P. M. Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *J. Geol.* **100,** 525–544 (1992).
37. Graham, S. T., Famiglietti, J. S. & Maidment, D. R. Five-minute, 1/2 degrees, and 1 degrees data sets of continental watersheds and river networks for use in regional and global hydrologic and climate system modeling studies. *Wat. Resour. Res.* **35,** 583–587 (1999).
38. Syvitski, J. P. M., Vorosmarty, C. J., Kettner, A. J. & Green, P. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. *Science* **308,** 376–380 (2005).
39. Rosenheim, B. E. *et al.* River discharge influences on particulate organic carbon age structure in the Mississippi/Atchafalaya River System. *Glob. Biogeochem. Cycles* **27,** 154–166 (2013).
40. Gaillardet, J., Dupré, B., Louvat, P. & Allègre, C. J. Global silicate weathering and CO_2 consumption rates deduced from the chemistry of large rivers. *Chem. Geol.* **159,** 3–30 (1999).



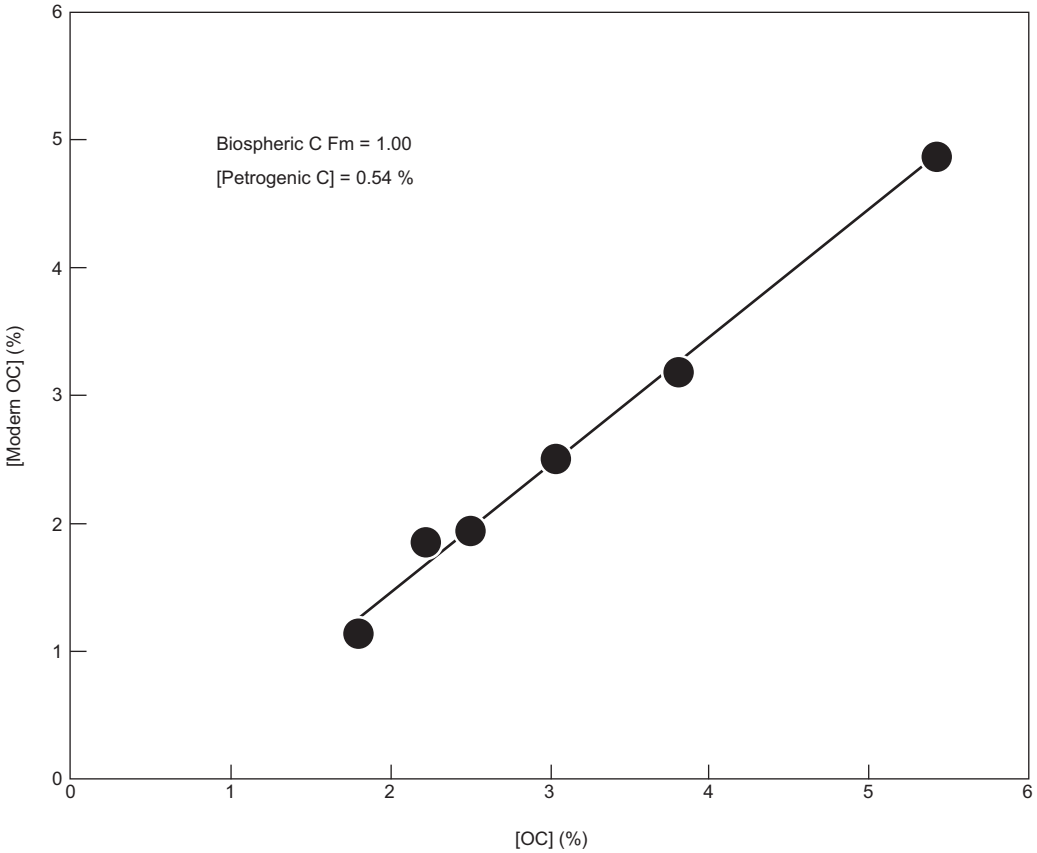
Extended Data Figure 1 | Global relationship between biospheric POC yield and NPP. Basin-averaged NPP estimates were derived from the MOD17 database[24].



Extended Data Figure 2 | Global relationship between long-term CO_2 sequestration yield and sediment yield. CO_2 sequestration through terrestrial biospheric POC burial (black dots; Y_{Corg}) is more sensitive to sediment yield than CO_2 sequestration through silicate weathering (grey crosses; Y_{Csil}). At high physical erosion rates (that is, high sediment yield), the burial of terrestrial biospheric POC becomes the dominant long-term atmospheric CO_2 sink. The dotted lines show CO_2 sequestration through terrestrial biospheric POC burial for the entire set of biospheric POC export data (Fig. 2), assuming constant burial efficiencies (BE) of 30 and 100%. CO_2 sequestration data through silicate weathering are from Gaillardet *et al.*[40]. $P = 0.001$; $r^2 = 0.80$.



Extended Data Figure 3 | Organic carbon and radiocarbon contents of bulk suspended sediments and grain size fractions in the Mississippi River. Results are expressed as modern organic carbon (that is, the product of modern fraction (Fm) and organic carbon content). The linear best fit gives the absolute petrogenic OC content (0.05%) as well as the Fm of the biospheric POC (0.88). Data from Wakeham *et al.*[31] and Rosenheim *et al.*[39]. $P = 0.001$; $r^2 = 0.99$.



Extended Data Figure 4 | Organic carbon and radiocarbon contents of bulk suspended sediments from the Ishikari River, collected over a wide range of flow regimes. Results are expressed as in Extended Data Fig. 3. The linear best fit gives the absolute petrogenic OC content (0.54%) as well as the Fm of the biospheric POC (1.00). Data from Alam *et al.*[32]. $P = 0.001$; $r^2 = 0.99$.

LETTER

doi:10.1038/nature14238

Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes

Jenni Hultman[1]†, Mark P. Waldrop[2], Rachel Mackelprang[3,4], Maude M. David[1], Jack McFarland[2], Steven J. Blazewicz[2], Jennifer Harden[2], Merritt R. Turetsky[5], A. David McGuire[6], Manesh B. Shah[7]†, Nathan C. VerBerkmoes[7], Lang Ho Lee[8], Kostas Mavrommatis[4]† & Janet K. Jansson[1,4,9,10]†

Over 20% of Earth's terrestrial surface is underlain by permafrost with vast stores of carbon that, once thawed, may represent the largest future transfer of carbon from the biosphere to the atmosphere[1]. This process is largely dependent on microbial responses, but we know little about microbial activity in intact, let alone in thawing, permafrost. Molecular approaches have recently revealed the identities and functional gene composition of microorganisms in some permafrost soils[2–4] and a rapid shift in functional gene composition during short-term thaw experiments[3]. However, the fate of permafrost carbon depends on climatic, hydrological and microbial responses to thaw at decadal scales[5,6]. Here we use the combination of several molecular 'omics' approaches to determine the phylogenetic composition of the microbial communities, including several draft genomes of novel species, their functional potential and activity in soils representing different states of thaw: intact permafrost, seasonally thawed active layer and thermokarst bog. The multi-omics strategy reveals a good correlation of process rates to omics data for dominant processes, such as methanogenesis in the bog, as well as novel survival strategies for potentially active microbes in permafrost.

We collected replicate permafrost soil cores (including the seasonally thawed active layer) from a permafrost plateau and an adjacent young thermokarst bog from a site in interior Alaska that represents future anticipated ecosystem transformations as permafrost thaws[7]. Owing to the recognized difficulty in cultivating the majority of microorganisms from soil[8], we combined several culture-independent 'omics' approaches: targeted 16S rRNA gene sequencing (16S) to determine the microbial community composition; total metagenomic DNA sequencing (MG) to determine the complement of phylogenetic and functional genes; total metatranscriptome RNA sequencing (MT) to determine which genes were expressed; and shotgun mass spectrometry-based metaproteomics (MP) to determine which proteins were produced. Together these analyses resulted in a large amount of data including 84.2 gigabases (Gb) of MG sequence, 20.4 Gb of MT sequence and approximately 7,000 proteins, which are among the highest yields obtained for any soil type so far[9,10]. We also compared the different molecular approaches among each other and to measured process rates. We analysed relative differences in gene expression among samples on the basis of the ratio of functional gene transcripts to genes (MT/MG).

The permafrost, active layer and thermokarst bog soils had a unique complement of genes, transcripts and proteins (Extended Data Fig. 1). Bacteria and Archaea dominated the data, with only a small fraction corresponding to fungi (MG 0.07 ± 0.04%, MT 0.41 ± 0.47%, zero in MP), so we omitted fungi from further analyses. There was greater overlap in shared genes between the soils, compared with transcripts and proteins (Extended Data Figs 1 and 2). Permafrost had fewer and more unique transcripts and proteins than active layer and thermokarst bog soils (Extended Data Figs 1 and 2). By contrast, many transcripts were shared between the bog and active layer, including many transporters (Extended Data Figs 1d and 3).

The most abundant bacteria found in the permafrost soil on the basis of 16S were members of Chloroflexi (19.2 ± 6.3% s.d.), Proteobacteria (17.9 ± 1.3%) and Actinobacteria (12.5 ± 3.1%) (Fig. 1a). Similar distributions were observed in 16S, MG and MT data sets (correlation coefficient $r > 0.9$ for 16S:MG and MG:MT; $r > 0.8$ for 16S:MT, Fig. 1b). However, some differences were found, such as higher representation of Actinobacteria in MG than 16S (27% and 12%, respectively; $P = 0.007$), partly explained by biases used in the 16S approach. Also, some phyla were more or less represented in the MT than the 16S and MG data, and we propose that these differences can be used to predict their relative activity at the time the samples were collected. In permafrost, the MT/MG ratios were highest for Proteobacteria (MT/MG = 1.7), Acidobacteria (MT/MG = 1.4), and Firmicutes (MT/MG = 1.6), suggesting that these representatives were acclimated to be active in subzero temperatures (Fig. 1b). Also, binning of MG assemblies from permafrost resulted in a draft genome of a novel *Acidobacterium* (PF_400, Extended Data Table 1) with closest sequence similarity to iron-reducing *Acidimicrobium ferrooxidans* (Genome-to-Genome Distance Calculator[11] DNA–DNA hybridization (DDH) 19.60% ± 2.30). The draft genome contained many interesting features including genes involved in Fe(III) transport, and cytochromes putatively involved in iron uptake and reduction (Extended Data Table 1). The genetic capacity for iron reduction in permafrost correlated well with dissimilatory Fe(III) reduction rates measured at the site (Table 1).

The active layer exhibited both more species and functional diversity than the other two soils (Fig. 1a), consistent with previous findings[2,3,12]. Members of the Actinobacteria, Acidobacteria and Proteobacteria phyla had the highest MT/MG ratios (Fig. 1b), suggesting that they were among the most active members in the seasonally thawed soil. A draft genome of a novel member of one of the Proteobacteria (AL_334) was also obtained from the active layer MG (Extended Data Table 1). Functional screening of the MG and MT data revealed many genes involved in specific biogeochemical cycles, including those for denitrification, nitrate reduction, iron reduction and methane oxidation (see Supplementary Information for brief overview), correlating with process measurements at the site (Table 1).

[1]Earth Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California, 94720, USA. [2]US Geological Survey, 345 Middlefield Road, Menlo Park, California 94025, USA. [3]Biology Department, 18111 Nordhoff Street, California State University Northridge, Northridge, California 91330, USA. [4]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [5]Department of Integrative Biology, 50 Stone Road East, University of Guelph, Guelph, Ontario N1G 2W1, Canada. [6]US Geological Survey, Alaska Cooperative Fish and Wildlife Research Unit, 211A Irving I Building, University of Alaska Fairbanks, Fairbanks, Alaska 99775, USA. [7]Chemical Sciences Division, One Bethel Valley Road, Building 1059, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6420, USA. [8]Graduate School of Genome Science and Technology, University of Tennessee and Oak Ridge National Laboratory, 2510 River Drive, Knoxville, Tennessee 37996, USA. [9]Department of Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, Berkeley, California 94720, USA. [10]Center for Permafrost Research (CENPERM), Department of Biology, Universitetsparken 15, University of Copenhagen, Copenhagen, DK-2100 Copenhagen, Denmark. †Present addresses: Department of Food Safety and Environmental Health, Agnes Sjöbergin katu 2, University of Helsinki, Helsinki 00014, Finland (J.H.); Oak Ridge National Laboratory, Biosciences Division, Tennessee 37831, USA (M.B.S.); Celgene Corporation, 1500 Owens Street, San Francisco, California 94158-2335, USA (K.M.); Biological Sciences Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99352, USA (J.K.J.).

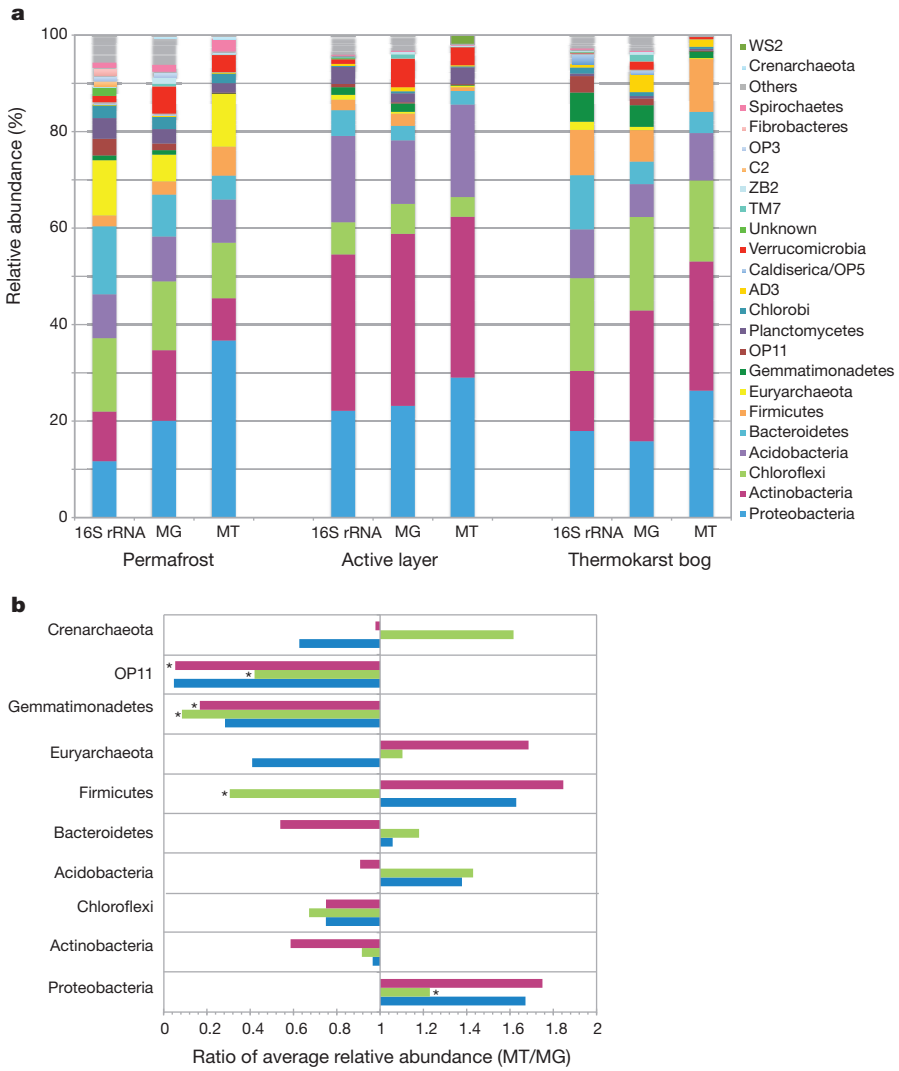


Figure 1 | Microbial phylogenetic composition in permafrost, active layer and thermokarst bog soils. **a**, Relative abundance of phyla on the basis of targeted 16S rRNA gene sequencing (16S), 16S rRNA gene reads from metagenomes (MG) and from metatranscriptomes (MT). The data represent the average of four replicates for 16S, two for MG and two for MT, except for permafrost MT which was obtained from only one sample owing to low RNA yield. **b**, Ratios of average relative abundance of ten abundant phyla in the metatranscriptomes relative to the metagenomes (MT/MG). Blue, permafrost; green, active layer; pink, bog. Bars marked with an asterisk are significantly different (two-tailed *t*-test, a *P* value of 0.05 was used as the significance level) from MT/MG = 1.0, to highlight potentially relatively higher activity.

In the thermokarst bog, the combined omics data correctly predicted that methanogenesis was predominant where high rates of methane production were observed (Table 1 and Fig. 2). Several genes involved in methanogenesis (*mcrABG*) were detected in the MG and MT data sets. The MT/MG ratios for methanogenesis were higher in the bog than any other site (Fig. 2). Also, the Euryarchaea had high MT/MG ratios of 1.69 ± 1.14 (Fig. 1b), and methanogen 16S rRNA gene sequences, transcripts and proteins were highly abundant (Figs 1–3). Of all the 16S sequences, 6.8–10.5% were assigned to *Methanosarcina*, a metabolically diverse methanogen that can produce methane using a variety of metabolic routes, including the acetoclastic pathway (Fig. 1a and Extended Data Fig. 4). We also binned three draft methanogen genomes from the bog MG (Extended Data Table 1), with the largest bin, Bog_15, containing genes for methane production via the hydrogenotrophic route, several of which were expressed and detected in the MT and the MP as well (Extended Data Table 1). A methanogen genome binned from an arctic wetland MG, *Methanoflorens stordalenmirensis*, was recently reported[4], with low sequence similarity to those we found here (DDH

Table 1 | Soil characteristics and process rates from the permafrost, active layer and thermokarst bog

	Permafrost	Active layer	Bog
Plant community		Black spruce/feathermoss	Sphagnum moss
Soil dissolved organic carbon (mg g^{-1})	0.08 ± 0.01†	1.05 ± 0.18*	0.95 ± 0.21*
Soil total dissolved nitrogen (mg g^{-1})	0.02 ± 0.002†	0.08 ± 0.01*	0.07 ± 0.01*
pH	5.76 ± 0.16*	4.69 ± 0.10†	4.88 ± 0.02†
Aerobic respiration (μg C h^{-1} g^{-1})	0.09 ± 0.01†	0.09 ± 0.01†	1.61 ± 0.40*
Anaerobic respiration (μg C h^{-1} g^{-1})	0.02 ± 0.002†	0.03 ± 0.01†	1.21 ± 0.27*
Methanogenesis (ng C h^{-1} g^{-1})	ND	ND	454 ± 150
Denitrification (ng N_2O h^{-1} g^{-1})	1.0 ± 0.7†	3.0 ± 0.8*†	6.4 ± 1.4*
Nitrate reduction (ng NO_3 h^{-1} g^{-1})	0.19 ± 0.10*	0.40 ± 0.08*	132 ± 36†
Fe reduction (ng Fe h^{-1} g^{-1})	10.0 ± 8.0†	26.2 ± 3.1*	ND
Sulphate reduction (ng SO_4 h^{-1} g^{-1})	2.9 ± 0.7†	4.2 ± 0.6*†	71 ± 17*
Aerobic methane oxidation (ng C h^{-1} g^{-1})	NA	200 ± 150	700 ± 400
Moisture (%)	58.2 ± 3.3†	81.4 ± 2.7*	83.5 ± 1.2*
Electrical conductivity (μSiemens)	379 ± 53*	272 ± 19*	242 ± 13*

Units are means ± 1 s.e.m. ND, not detected; NA, not determined. Footnote symbols (*, †) denote significant differences among soils ($P < 0.05$).

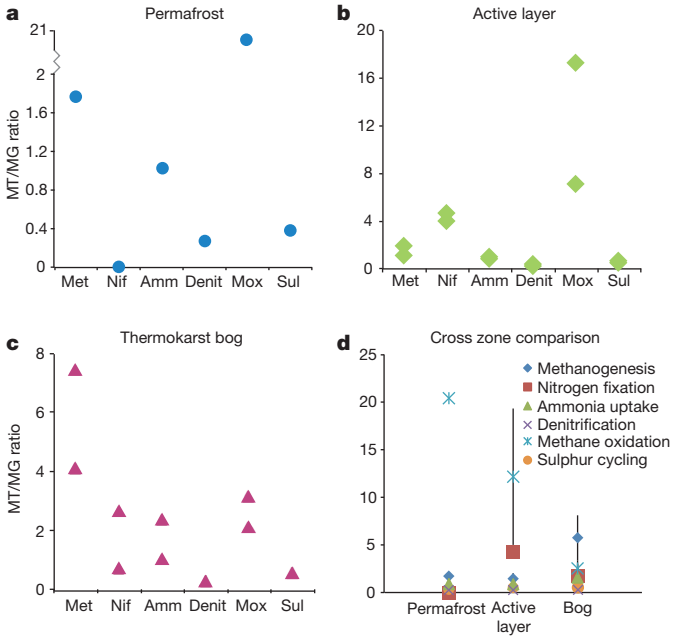


Figure 2 | Average MT/MG ratios. **a**, Permafrost; **b**, active layer; **c**, thermokarst bog soils for two replicate cores. **d**, Comparison of MT/MG on the same plot. Error bars, s.e.m. Amm, ammonia uptake; Denit, denitrification; Met, methanogenesis; Mox, methane oxidation; Nif, nitrogen fixation; Sul, sulphate reduction.

20.60–37.30%). Closest sequenced relatives to our draft genomes were *Methanosarcina acetivorans* (Bog_301, 69.70% ± 2.92) and *Methanosaeta concilii* (with Bog_232 46.60% ± 2.58 and Bog_15 42.70 ± 2.53 DDH similarity (Extended Data Table 1)); thus they represent previously undescribed methanogen genomes.

Two novel Chloroflexi draft genomes were also binned from the bog MG (Extended Data Table 1). One of them, Bog_440, with closest similarity to a representative of the genus *Anaerolinea* (*A. thermophila*, DDH 15.20% ± 2.14), was phylogenetically similar to a dominant operational taxonomic unit (OTU) in the 16S data (Fig. 1a and SOGA31 in Extended Data Fig. 5) and has previously been reported as abundant in other arctic samples on the basis of 16S[3,13], but represents an undersampled phylum with few isolates and none with sequence similarity to those we found here. Draft Bog_440 possessed some genes involved in sulphate reduction and iron utilization (Extended Data Table 1), suggesting that it may play a role in these processes.

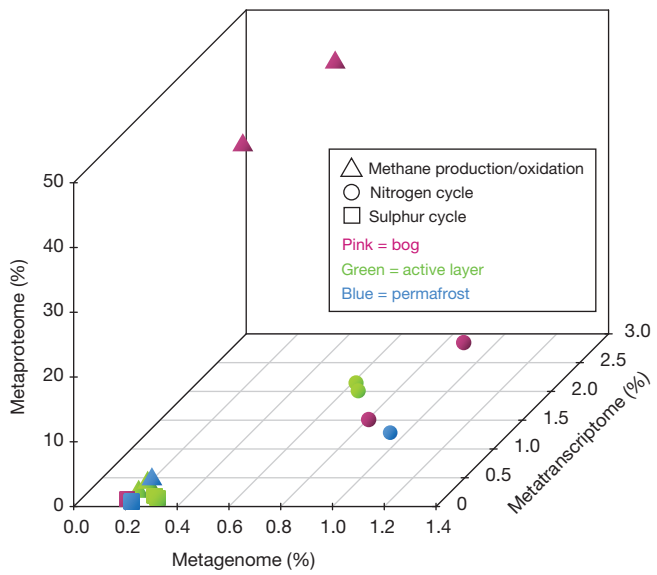


Figure 3 | Relationships among MG, MT and MP for nitrogen cycling (circle), sulphur cycling (square) and methanogenesis (triangle) in permafrost (blue), active layer (green) and thermokarst bog soils (pink). Points represent the sum of all data involved in the particular process.

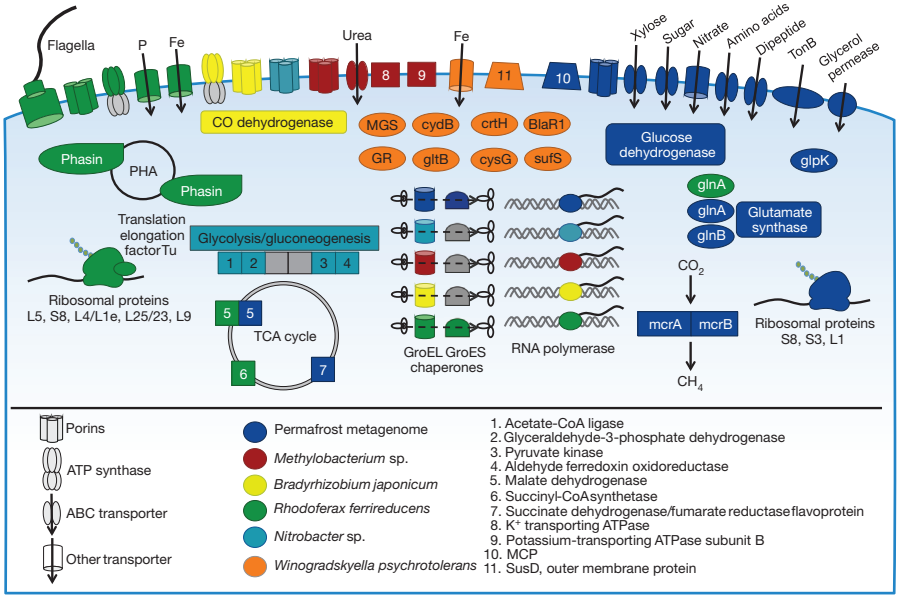
The metaproteomics data enabled us to detect the dominant microbial proteins actually produced and/or stabilized in the soil. Most of the proteins matched to the metagenomes and many were 'uncharacterized' or 'hypothetical' (Table 2 and Supplementary Table 1), or predicted to be involved in housekeeping functions (Extended Data Fig. 1 and Fig. 4). In permafrost, we detected a relatively high representation of cold-shock proteins, presumably for survival under frozen conditions (Fig. 4a and Supplementary Table 1). Surprisingly, proteins involved in chemotaxis and motility were also observed in permafrost (Supplementary Table 1), suggesting that some members of the community had the capacity for motility. Fewer transporters were detected in permafrost than active layer and bog soils (Fig. 4a and Supplementary Table 1), as also reflected by lower levels of transporter transcripts in permafrost (Extended Data Supplementary Fig. 3). Several cold-shock proteins were also detected in the active layer (Fig. 4b), reflecting the need for the community to cope with not only the seasonal shifts in temperatures, but also exposure to extreme freezing conditions. Other abundant proteins in the active layer included several transporters (Extended Data Fig. 1, Fig. 4b and Supplementary Table 1), indicative of greater potential for microbial transport of nutrients in the thawed soils (Extended Data Figs 1b, d and 3). Unlike the permafrost and active layer, there were fewer proteins linked to cold tolerance and other stress-related reactions in the bog soils (7%, 14% and 12% in bog, active layer and permafrost, respectively) (Supplementary Table 1).

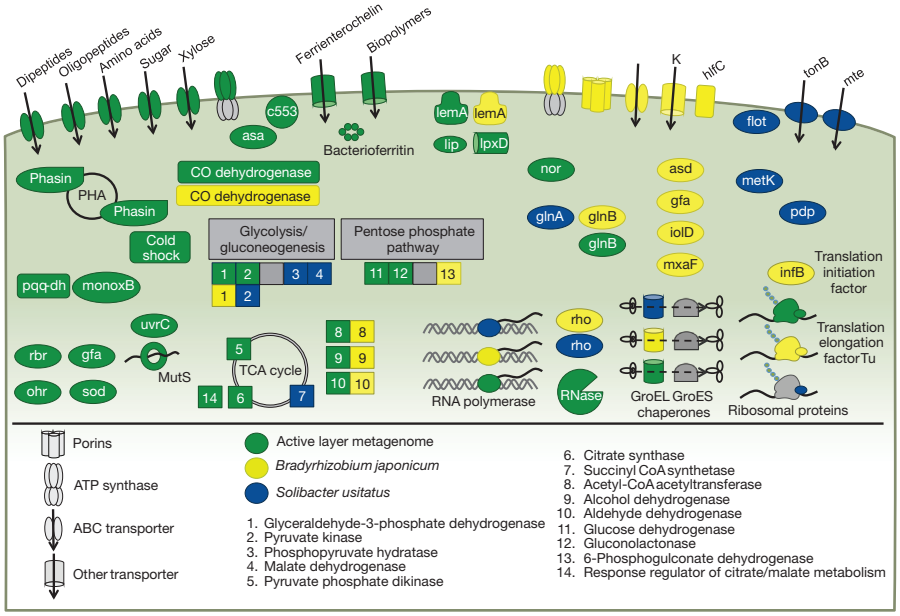
In addition to proteome searches against the matched metagenomes, we also performed searches to microbial genomes: 180 environmental isolates, plus an additional 13 recently sequenced isolates from cold environments (Supplementary Table 3). *Rhodoferax ferrireducens* had many (58) protein matches in permafrost (Table 2 and Supplementary Table 1) although it was relatively rare in the MT (<0.001%), MG (<0.005%) and 16S sequences (0.1%). Several proteins were also assigned to some betaproteobacteria, nitrifiers, acidobacteria, facultative methylotrophs (Fig. 4a and Supplementary Table 1) and to several psychrophile isolates (Fig. 4a and Extended Data Table 2). More proteins were assigned to a wider range of the reference genomes in the active layer than permafrost soil (Table 2), indicating a larger diversity of active species in this soil, including several Proteobacteria, in addition to 11 of

Table 2 | Twenty most abundant hits of metaproteome data to genomes/metagenomes

Reference metagenome/genome	Permafrost	Active layer	Bog
Permafrost metagenome	**41.6 ± 3.3**	**32.3 ± 7.4**	**21.1 ± 12.3**
Active layer metagenome	**54.6 ± 37.4**	**50.5 ± 8.2**	**21.8 ± 22.5**
Bog metagenome	**38.4 ± 0.3**	**46.6 ± 4.8**	**73.3 ± 10.7**
Bradyrhizobium japonicum USDA 110	**2.6 ± 1.1**	**6.3 ± 2.1**	**2.6 ± 2.1**
Burkholderia xenovorans LB400	**2 ± 2.3**	**3.0 ± 2.1**	0.1 ± 0.1
Arthroderma benhamiae CBS 112371	0.3 ± 0.4	**2.8 ± 3.2**	**2.3 ± 2.7**
Rhodopseudomonas palustris CGA009	1.0 ± 0.3	**2.4 ± 1.4**	**1.5 ± 1.5**
Nitrobacter winogradskyi Nb-255	**1.8 ± 0.8**	**2.2 ± 0.9**	1.0 ± 0.9
Burkholderia cepacia 383	1.5 ± 1.7	**2.0 ± 1.1**	0.1 ± 0.1
Delftia acidovorans SPH-1	1.4 ± 1.9	**1.9 ± 1.7**	0.4 ± 0.6
Methanosaeta concilii GP-6	0.1 ± 0.0	0.0 ± 0.0	**2.7 ± 2.3**
Methanosarcina mazei Go1, DSM 3647	0.0 ± 0.0	0.0 ± 0.0	**2.0 ± 0.4**
Solibacter usitatus Ellin6076	1.3 ± 1.1	1.6 ± 0.9	**1.5 ± 1.0**
Nitrobacter sp. Nb-311A	0.9 ± 0.2	1.7 ± 1.3	**1.1 ± 0.9**
Rhodoferax ferrireducens T118	**4.0 ± 2.6**	1.7 ± 0.9	0.5 ± 0.6
Acidovorax sp. JS42	**1.9 ± 1.6**	1.6 ± 1.0	0.4 ± 0.5
Ensifer medicae WSM419	**1.9 ± 2.5**	0.4 ± 0.4	0.4 ± 0.0
Ralstonia solanacearum GMI1000	**1.8 ± 1.6**	1.8 ± 0.8	0.3 ± 0.2
Methylobacterium nodulans ORS 2060	1.7 ± 1.2	1.2 ± 0.2	1.0 ± 0.3

The ten most abundant database hits are highlighted in bold for each soil. Each metagenome data set (first three rows) was obtained from the same source sample as the metaproteomes. Based on the average normalized spectral abundance factor (NSAF) values and standard deviation of four replicate Velos/Orbitrap mass spectrometer runs for each soil.





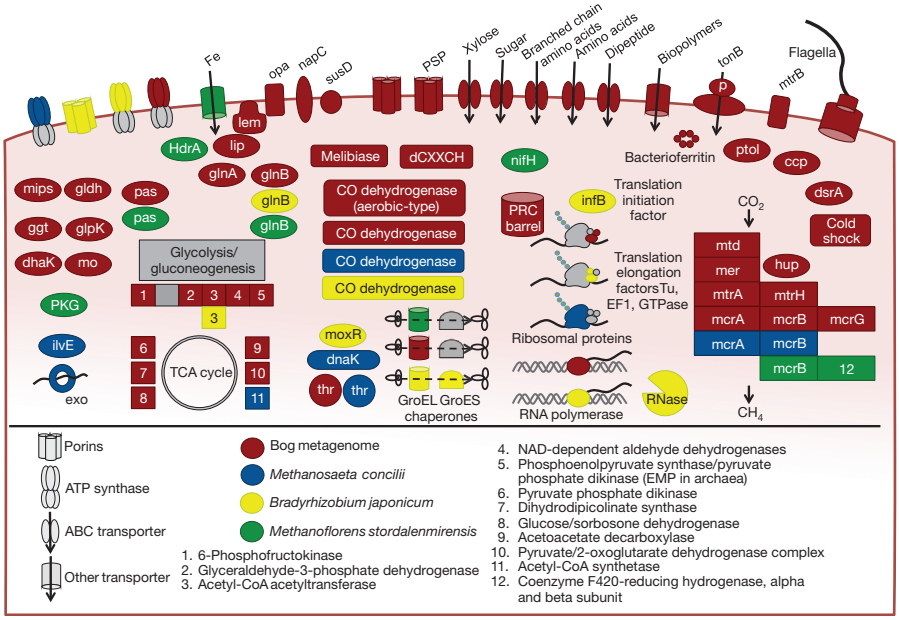


Figure 4 | Visualization of proteins identified from metaproteomics data sets. **a**, Permafrost; **b**, active layer; **c**, thermokarst bog soils. The proteins were identified by searching against databases comprising matched metagenomes from each site and for genomes from selected sequenced microbial isolates. Grey shading indicates predicted proteins that were not detected. Key to abbreviations: asa, arylsulphatase A; asd, aspartate-semialdehyde dehydrogenase; BlaR1, regulatory sensor-transducer, BlaR1/MecR1 family; c553, cytochrome *c*-553; ccp, cytochrome *c* peroxidase; crtH, carotenoid *cis*-transisomerase; cydB, cytochrome *d* ubiquinol oxidase subunit I; cysG, uroporphyrinogen-III methyltransferase; dcxxch, doubled CXXCH domain containing protein; dhaK, dihydroxyacetone kinase; dnaK, DnaK chaperone protein; dsrA, sulphite reductase subunit A; exo, exosome complex exonuclease 1; flot, flotillin protein; gfa, *S*-(hydroxymethyl) glutathione synthase; ggt, γ-glutamyltransferase; gldh, glutamate dehydrogenase; glnA, glutamate synthetase; glnB, nitrogen regulatory protein P-II; glpK, glycerol kinase; gltB, ferredoxin-dependent glutamate synthase; GR, glutathione reductase-like proteins; HdrA, heterodisulphide reductase subunit A and related polyferredoxins; hflC, hydrolase serine protease transmembrane protein; hup, heterosulphide reductase; ilvE, branched-chain amino-acid aminotransferase; infB, translation initiation factor IF-2; iolD, malonic semialdehyde oxidative decarboxylase; lem, peptidoglycan-associated lipoprotein dehydrogenase; lemA, outer membrane protein and related peptidoglycan-associated; (lipo) proteins; lip, periplasmic or secreted lipoprotein; lpxD, UDP-3-O-(3-hydroxymyristoyl) glucosamine *N*-acetyl transferase; mer, 5,10-methylenetetrahydromethanopterin reductase; metK, methionine adenosyltransferase; MGS, methylglyoxal synthase; mips, myo-inositol-1-phosphate synthase; mo, methane monooxygenase/ammonia monooxygenase; monoxB, Monooxygenase subunit B protein; moxR,MoxR family protein; Mtd, coenzyme F420-dependent N(5),N(10)-; methenyltetrahydromethanopterin; mte, MotA/TolQ/ExbB proton channel; mtrA, N5-methyltetrahydromethanopterin,coenzyme M; methyltransferase subunit A; mtrB, decahaem-associated outer membrane protein; mtrH, tetrahydromethanopterin *S*-methyltransferase, subunit H; mutS, mismatch repair ATPase (MutS family); mxaF, methanol dehydrogenase; napC, NapC/Nir T cytochrome *c* family; nor, nitrous oxide accessory protein; ohr, peroxiredoxin, Ohr subfamily; opa, opacity protein; p, TonB-dependent receptor plug; pas, PAS domain S box; pdp, pyrimidine-nucleoside phosphorylase; PKG, phosphoglycerate kinase; pqq-dh, PQQ-dependent dehydrogenase methanol/ethanol family; PSP, phosphate-selective porin; ptol, periplasmic component of the Tol biopolymer transport system; rbr, rubrerythrin; rho, transcription termination factor Rho; RNase, ribonuclease E; rng, ribonuclease G; sod, superoxide dismutase; thr, thermosome; uvrC, exinuclease ABC, C subunit.

the *R. ferrireducens* proteins mentioned above (Supplementary Table 1), and to some of the psychrophile isolates with the majority matching to *Rhodonellum psychrophilum* and *Winogradskyella psychrotolerans* (Supplementary Table 3). In the bog, several proteins matched to methanogens, including *Methanosarcina mazei*, *Methanosaeta concilii*, *Methylobacterium nodulans*, *Methanoflorens stordalenmirensis* (Table 2, Supplementary Table 3 and Fig. 4c) and to our draft methanogen genomes, reflecting the high rates of methanogenesis in the bog. These included proteins involved in methanogenesis, nitrogen metabolism, iron transport and tolerance of cold and redox induced stress (Fig. 4c). As in the other samples, many proteins also matched to some of the psychrophile genomes (Supplementary Table 3).

Comparison of the MG, MT and MP data from the three soils provides insights into the linkages between omics data and processes. Processes that were detected in each of the three omics data sets, namely nitrogen, sulphur and methane cycling pathways, are shown in Fig. 3 (see Supplementary Information for details). The results indicate that permafrost microbial communities have a lower general functional potential than thawed soils, but also provide insights into adaptation strategies for life in frozen conditions. We found that global molecular data were a poorer predictor of biogeochemical process rates in permafrost than the other soils, probably owing, in part, to the frozen state of the soil. All measured process rates were generally low in frozen permafrost, but relative abundances of transcripts and proteins for several genes and proteins of some specific pathways, such as methane oxidation, were nonetheless similar to those for the active layer. It is possible that some of the RNA and proteins were preserved for long periods in this frozen environment, or that they were maintained in microbes that are surviving in a dormant state. Alternatively, some microbes in permafrost are actively expressing these genes and translating them into proteins in subzero conditions. In particular, dissimilatory Fe(III) reduction emerged as a potential metabolic strategy used by microbes in permafrost, as supported by combined omics data and process rates. By comparison, the active layer communities expressed genes and proteins involved in obtaining energy and nutrients from a diversity of aerobic and anaerobic processes and were equipped with functions for survival under freeze–thaw conditions. The bog represented a different scenario with a very high measured rate of methanogenesis and correspondingly high relative abundances of genes, transcripts and proteins involved in methanogenesis, thus demonstrating the potential linkage between molecular data and ecosystem level process rates, particularly when the process is dominant.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 April 2014; accepted 19 January 2015.
Published online 4 March 2015.

1. Koven, C. D. *et al.* Permafrost carbon-climate feedbacks accelerate global warming. *Proc. Natl Acad. Sci. USA* **108,** 14769–14774 (2011).
2. Yergeau, E., Hogues, H., Whyte, L. G. & Greer, C. W. The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. *ISME J.* **4,** 1206–1214 (2010).
3. Mackelprang, R. *et al.* Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480,** 368–371 (2011).
4. Mondav, R. *et al.* Discovery of a novel methanogen prevalent in thawing permafrost. *Nat. Commun.* **5,** 3212 (2014).
5. Schuur, E. A. & Abbott, B. Climate change: high risk of permafrost thaw. *Nature* **480,** 32–33 (2011).
6. McCalley, C. K. *et al.* Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514,** 478–481 (2014).
7. Jorgenson, M. T., Racine, C. H., Walters, J. C. & Osterkamp, T. E. Permafrost degradation and ecological changes associated with a warming climate in central Alaska. *Clim. Change* **48,** 551–579 (2001).
8. Jansson, J. K. Towards "Tera-Terra": terabase sequencing of terrestrial metagenomes. *Microbe* **6,** 309–315 (2012).
9. Chourey, K. *et al.* Direct cellular lysis/protein extraction protocol for soil metaproteomics. *J. Proteome Res.* **9,** 6615–6622 (2010).
10. Nicora, C. D. *et al.* Amino acid treatment enhances protein recovery from sediment and soils for metaproteomic studies. *Proteomics* **13,** 2776–2785 (2013).
11. Auch, A. F., von Jan, M., Klenk, H. P. & Goker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* **2,** 117–134 (2010).
12. Waldrop, M. P. *et al.* Molecular investigations into a globally important carbon pool: permafrost-protected carbon in Alaskan soils. *Glob. Change Biol.* **16,** 2543–2554 (2010).
13. Liebner, S., Harder, J. & Wagner, D. Bacterial diversity and community structure in polygonal tundra soils from Samoylov Island, Lena Delta, Siberia. *Int. Microbiol.* **11,** 195–202 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Hettich and the Organic and Biological Mass Spectrometry group at Oak Ridge National Laboratory for access to mass spectrometry instrumentation. M. Haw, K. Li, K. Chavarria and R. Lamendella are acknowledged for help with pre-processing frozen samples. We thank K. Billis for help with RNA sequence preprocessing. This work was partly supported by the Director, Office of Science, Office of Biological and Environmental Research, Climate and Environmental Science Division, of the US Department of Energy, Terrestrial Ecosystem Science-Scientific Focus Area (TES-SFA), through a Community Sequencing Project at the DOE Joint Genome Institute (JGI CSP - 152) and by a Lawrence Berkeley National Laboratory Laboratory Directed Research & Development (LDRD) grant, all under contract number DE-AC02-05CH11231; by the Pacific Northwest National Laboratory under contract number DE-AC05-76RL01830; and by the Danish National Research Foundation (CENPERM DNRF100). The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231. Additional funding and considerable logistic support were provided by the Bonanza Creek Long-Term Ecological Research Program, which is jointly funded by National Science Foundation (DEB 1026415) and the US Department of Agriculture Forest Service, Pacific Northwest Research Station (PNW01-JV112619320-16). Support was also received from the US Geological Survey Climate R&D Program and Alaska Climate Science Center. J.Hu. was supported by Academy of Finland grant number 135669.

Author Contributions J.K.J., M.P.W. and J.Hu. planned the study. M.P.W. and J.M. collected the samples and J.M. performed the chemical analyses. J.Ha., M.R.T. and A.D.M. performed the site characterization. M.B.S., N.C.V., M.M.D. and L.H.L. performed the proteomics. J.Hu., R.M., S.J.B. and K.M. analysed the sequence data. J.Hu., M.P.W. and J.K.J. wrote the paper with contributions from all authors. J.Hu., J.K.J., M.P.W. and R.M. made the figures.

Author Information The 454 nucleotide sequences have been deposited in Sequence Read Archive under BioProject number PRJNA222786, metagenomes and metatranscriptomes in IMG/M under Study ID Gs0063124 MG-RAST under project number 11953 and proteomics data in the PRIDE partner repository under dataset identifier PXD001131. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.K.J. (janet.jansson@pnnl.gov).

METHODS

No statistical methods were used to predetermine sample size.

Site description. The sites used in this experiment are part of the Alaska Peatland Experiment (APEX), which is part of the Bonanza Creek Long-Term Ecological Research Program just outside Fairbanks, Alaska (64.70° N, −148.3° W). The forested site is underlain by permafrost, is dominated by black spruce (*Picea mariana*) and has an understorey composed of *Vaccinium uliginosum*, *Vaccinium vitis-idaea*, *Calamagrostis canadensis* and occasionally *Eriophorum vaginatum*, *Larix laricina* and *Salix bebbiana*. Understorey non-vascular vegetation is composed of moss species *Hylocomium splendens*, *Sphagnum* sp., *Dicramium* sp. and lichen.

The thermokarst bog site is primarily composed of *Sphagnum* mosses, as well as vascular species *Carex aquatilis*, *Eriophorum chamissonis*, *Chamaedaphne calyculata* and *Carex chordorrhiza*, and contains recently deceased black spruce trees.

During the time of sample collection, the temperature of the surface permafrost was just below freezing, much like most of the surface permafrost in interior Alaska (60 cm: maximum −0.052 °C, minimum −2.18 °C). In the thermokarst bog, the soil temperature ranged from 3.49 to −0.08 °C at 100 cm to 6.6 to −2.69 °C at 50 cm. The pH at the sites was 4.8 in active layer, 4.9 and 5.8 in permafrost (1:1 soil water ratio).

Sample collection. Samples of the active layer and permafrost (black spruce site) were taken in summer 2009 (for DNA extraction, four cores per soil) and summer 2012 (for process rate measurements) by using a SIPRE corer that produces a 7.5 cm diameter core down to 1 m. Maximum active layer depth was measured at 55 cm at this site. Samples of the active layer were at approximately 30–35 cm depth and of permafrost were 65–75 cm depth.

The thermokarst site underwent a thaw approximately 100 years ago supported by air photographic interpretation (Turetsky) and radiocarbon analysis of the transition between permafrost and thermokarst soils. Cores (for DNA extraction, four cores) were taken using a sharpened polyvinyl chloride (PVC) pipe of diameter 4 feet (~1.22 m) and by cutting around the bottom of the PVC pipe as it was slowly inserted into the bog, without compacting the organic materials. When the PVC pipe was inserted as far as it could go (approximately 1.2 m), the bottom of the PVC tube was covered with a PVC cap and the core was slowly raised out of the ground. The PVC core was quickly wrapped in a plastic bag, taped tightly and placed into a cooler with dry ice to be immediately frozen. Samples for microbial analysis were taken at approximately 100 cm depth. Samples for process rate measurements were collected in April 2011 using a SIPRE corer. In the field, cores were wrapped in aluminium foil and packed in dry ice for transport to our laboratory facilities, where they were kept frozen pending subsampling. Process measurements were conducted on core splits from about 45–60 cm below the surface.

Sample processing. For incubation studies, frozen soils were sectioned using a compound mitre saw. Before sectioning, cores were scraped of surface organic matter and other contaminants that froze, or otherwise adhered to the core during the drilling and extraction process. Frozen soil or peat aliquots were split from each core section for downstream analyses using a band saw to divide sections along the vertical profile. Core splits were thawed in a glove box under anaerobic conditions (1–2% H_2 atmosphere volume balanced with N_2) and homogenized by hand. Core splits were allocated either to standard soil analyses, for example moisture, pH and total CN content, or incubation units for sequential determination of Fe(III)/Fe(II), SO_4^{2-}, NO_3^-, denitrification potential, as well trace gas production (CO_2, CH_4 and N_2O) by aerobic and anaerobic pathways.

Soil analyses. Soil moisture was determined gravimetrically by drying samples to a constant mass at 65 °C. These subsamples were subsequently powdered in a modified roller mill and used to measure soil organic carbon and total nitrogen by combustion analysis on a LECO 2000 CNS Autoanalyzer calibrated with National Institute of Standards and Technology standards Buffalo River sediment and apple leaves (SRM 2704 and SRM 1515, respectively). Owing to the acidity of these samples, carbonate removal was not necessary before analysis. For gamma samples, soil pH was determined from a 1:2 (mass:volume) slurry of oven-dried soils and deionized water which was allowed to equilibrate at room temperature for 30 min. For beta samples, we measured pH on pore water collected in July using acid-washed (30 cm × 6.0 mm internal diameter), stainless steel sediment pore water samplers connected to 60 ml syringes fitted with three-way Luer-lock stopcocks. The samplers were flushed with several volumes of pore water before collecting and filtering (0.45 μm nylon) samples into sterile 50 ml polyethylene centrifuge tubes.

Rates of iron reduction, and sulphate and nitrate utilization. We established two series of six incubation units representing six sampling periods (0, 1, 3, 5, 8 and 12 d) for each of the forest and permafrost samples, and two series of four incubation units (0, 1, 5, 12 d) for each bog soil core. An additional incubation unit consisting of autoclaved soil was established for each bog core section, and harvested at 12 d. Incubation units consisted of about 5 g of homogenized whole soil or peat, subsampled under anaerobic conditions into 40 ml borosilicate glass vials fitted with 3.18 mm polypropylene septa (Environmental Sampling Supply). Each series was incubated in the dark at 5 or 22 °C under a microaerobic atmosphere maintained by a vacuum desiccator purged with five volumes of argon and held at a tension slightly below ambient pressure. On the appropriate sampling date, one vial from each series was destructively harvested by adding 15 ml of O_2-free deionized water to the vial and vortexing at maximum speed for 60 s. Pore water and solids were separated by centrifugation (15 min at 500*g*) and subsampled in the glove box to minimize potential for oxidation. Pore waters were filtered through a sterile 0.2 μm syringe filter into 15 ml polyethylene centrifuge tubes and stored frozen at −25 °C pending analysis for sulphate and nitrate content. Sulphate concentrations were determined using a Dionex DX ion chromatograph equipped with a 4 mm AS4A-SC column (Dionex Corporation). Peak areas for sulphate were quantified against a six-point calibration (0.5–20.0 μg ml^{-1}) developed from a commercial 1,000 μg ml^{-1} stock SO_4 solution (Ultra Scientific). Nitrate content was analysed colorimetrically using an automated discrete photometric analyser (Aquakem 250, Thermo Fisher Scientific) following United States Environmental Protection Agency method 353.1. Changes in SO_4 and NO_3 over time were used as estimates of reduction rates.

Rates of acid-extractable ferric iron (Fe(III)) reduction were estimated at multiple time points by measuring ferrous iron (Fe(II)) production colorimetrically using the ferrozine method[14]. Briefly, 0.5 g of homogenized whole soil, or peat, was transferred under anaerobic conditions into a 15 ml plastic centrifuge tube. We added 10 ml of 0.5 M HCl to the tube, and vortexed the mixture at maximium speed for 15 s. Tubes were shaken in the dark for 1 h at 200 revolutions per minute, and then centrifuged at 2,500*g* for 15 min. A 100 μl subsample of the supernatant was added to 5.0 ml of ferrozine solution (1.0 g l^{-1} 3-(2-pyridyl)-5,6-diphenyl-1,2,4-triazine-4′, 4″-disulphonic acid sodium in 50 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulphonic acid) adjusted to pH 4) and vortexed for 10 s. Absorbance was measured at 562 nm after allowing a full minute for colour development. To determine Fe(III), we subsequently added 0.25 ml hydroxylamine-HCl (0.25 M in 0.25 M HCl) to each tube, and re-read absorbance at 562 nm, after allowing 15 min for the reduction of Fe(III) to Fe(II). We used a six-point calibration of Fe(II) standards (2–80 μg ml^{-1}) to determine acid-extractable iron fractions. Amorphous ferric iron content was calculated as the difference between total acid-extractable iron and ferrous iron content before the addition of hydroxylamine. Fe(II) production over time was used as an estimate of reduction rate.

Denitrification rate measurements. We employed a modified acetylene inhibition technique[15] to evaluate rates of denitrification among soil and peat replicates. In an anaerobic glove box, we placed about 3.0 ml of material (soil or peat) under field-moist conditions into two series of six serum vials (13 ml volume), representing five sampling periods and a sterilized control, for each core section. Vials were crimp-sealed with butyl rubber stoppers, with the exception of sterilized controls, which were autoclaved (121 °C and 250 kPa) for 30 min, and allowed to cool in the glove box before sealing. Vials were amended with a 15% headspace addition of acetylene gas and inverted several times to improve diffusion of the acetylene. Each series of vials was incubated statically at 5 or 22 °C before sampling for N_2O. Sampling times were 10 min and 1, 2, 3 or 5 days after the addition of acetylene. At the appropriate sampling time, N_2O concentrations were measured by injecting 0.25 ml of headspace gas into an HNU model 301 gas chromatograph fitted with a Valco ^{63}Ni electron-capture detector and a 2.6 m (1.5 mm internal diameter) Porapak R 100/120 mesh column. Peak detection for N_2O was calibrated with certified N_2O-in-N_2 standards with concentrations of 0.101, 1.02 and 10.1 ppm by volume (Air Liquide America Speciality Gases). Denitrification rates were determined by calculating the change in N_2O over time.

Methane oxidation. Soils were collected from the bog and black spruce forest APEX sites in Bonanza Creek Long-Term Ecological Research on 18 June 2009. At each site, soil cores were taken every 10 m along 40 m transects with five soil cores (each core representing a replicate) collected along each transect. Bog and black spruce forest cores were collected down to the ice layer using a corer with a diameter of 4 feet. The living vegetation top layer was removed from each core. Bog cores were approximately 30 cm deep (measured from below the vegetative layer to the ice layer), and each core was split up into three depth samples: 0–10 cm, 10–20 cm and 20–30 cm. Only data from 20–30 cm are reported here. Forest cores were approximately 10 cm deep. Cores were kept on ice and transported to the University of California, Berkeley, where they were stored at 4 °C until incubation experiments were initiated.

CH_4 oxidation incubations. For CH_4 oxidation incubations, soils were delivered (10.0 ± 0.1 g) to sterile 250 ml jars. The jars were sealed with airtight lids containing septa to allow for headspace gas sampling. The initial headspace was composed of room air. To ensure that oxygen was not depleted during the CH_4 oxidation incubations, at the end of the incubations, headspaces were analysed for O_2 concentration, using a Probe YSI 52 oxygen meter. The final average headspace O_2 concentration was 19.2 ± 1.0%. For CH_4 production incubations, soils (10 g) were

loaded into 100 ml serum bottles, and bottles were sealed with gas-impermeable blue butyl rubber septa (Bellco). Anoxic conditions were created using a standard alternating vacuum and pressure approach (for example refs 16, 17). Briefly, bottle headspaces were evacuated by vacuuming to 51 kPa and then over-pressured with N_2 to 40 kPa, and this process was repeated ten times. On final filling, headspace was only filled to 13 kPa with N_2. The headspaces were then flushed with N_2/CO_2 (98/2%) at 135 kPa for 5 min through the septa using 23-gauge needles as input and exhaust ports. Headspace pressure was equilibrated to 10 kPa using a N_2 flushed syringe. For the anaerobic work, all transfers, additions and headspace samplings were done using gas-tight syringes fitted with stopcocks that were first degassed with N_2 to avoid O_2 contamination. For all anaerobic work, N_2 and N_2/CO_2 were first passed through hot copper fillings (~350 °C) to remove traces of O_2 (for example ref. 18), and the copper fillings were reduced daily using H_2. All incubations were setup with five replicates and with heat killed controls (autoclaved twice at 121 °C for 20 min). Methane (1,700 ppm) was delivered to all treatments. Soils were incubated for 5 days at 11 °C in the dark.

CH_4 oxidation gas analysis. Headspace samples (2 ml) were taken from incubation jars via gas tight syringe fitted with a stopcock at $t = 0$ (30 min after CH_4 was delivered), 24 h and 5 d. Headspace samples were transferred to 10 ml serum bottles sealed with thick black butyl rubber stoppers. Samples were analysed using an HP6890 GC (Hewlett Packard) fitted with a flame ionization detector and pulsed discharge ionization detector for CH_4 and CO_2 analysis, respectively. CH_4 oxidation and production rates were calculated by difference in total CH_4 concentration in the headspace at the beginning and end of the incubation period.

Laboratory incubation experiment and the rate of trace gas production. Trace gas (CH_4, CO_2, N_2O) production was monitored under aerobic and anaerobic conditions during a long-term (1 year) laboratory incubation experiment. We added 1–7 g (dry mass) of soil or peat, with field-moisture content, to 100 ml clear borosilicate glass serum vials (Sigma Aldrich) fitted with either 20 mm grey (aerobic) or blue (anaerobic) butyl rubber stoppers (Bellco Glass) and aluminium crimp seals. All vials were sealed and pre-incubated for 3 days in the dark before gas measurements were initiated. To prevent inhibition of respiration due to excessive CO_2 build-up, aerobic vials were capped for only 24 h before gas flux measurements, then aerated before being returned to their respective incubation chambers. Between sampling periods, these vials were covered with 0.8 mm polyethylene sheeting to prevent excessive moisture loss while still allowing gas exchange. Anaerobic vials were flushed with argon for 10 min after measurement to reset headspace atmosphere for the next sampling period. After gas sampling, the water content of all vials was checked gravimetrically and adjusted, as necessary, to the original field-moisture content with O_2-free deionized water. Incubations were conducted at 5 °C and room temperature (22 °C). We sampled the headspace of all vials weekly for 4 weeks. We removed 5 ml from the headspace of each vial using an air-tight syringe, fitted with a Luer-lock stopcock sealed with high-performance vacuum grease, and measured CO_2, CH_4 and N_2O with a gas chromatograph pre-configured for greenhouse gas monitoring (SRI 8610C gas chromatograph with flame ionization detector–methanizer and electron capture detector, column 3.2 mm internal diameter × 6 m HaySep D, 1 ml sample loop; SRI Instruments). Peak areas for trace gases were calibrated using a three-point curve (100, 1,000 and 10,000 ppm) for CO_2, a three-point curve (10, 100 and 1,000 ppm) for CH_4 and a two-point curve (0.101 and 1.02 ppm) for N_2O. Calibration curves were developed with certified gas standards balanced with N_2 (Air Liquide America Speciality Gases).

Statistical analysis for soil edaphic characteristics and process rates. One-way analysis of variance was used to compare moisture content, pH, electrical conductivity, dissolved organic carbon and total dissolved nitrogen among active layer, permafrost and bog soils ($n = 4$–5). In the cases where data were not normally distributed or could not be transformed to a normal distribution, non-parametric Wilcoxon tests were used. A P value of 0.05 was used as the significance level. For Fe(II) production, SO_4 loss, NO_3 loss, denitrification and CH_4 flux at 4 weeks under anaerobic conditions, aerobic CO_2 production at 4 weeks and anaerobic CO_2 production at 4 weeks, two-way analysis of variance was used with soil type and incubation temperature as the independent variables. In all cases, except for NO_3 loss, rates were higher with increased temperature. In Table 1, we present the main effect of 'soil type' and its mean and standard error. For each soil × temperature combination, $n = 5$.

Sample preparation for nucleic-acid extraction. To ensure that the samples were free of surface contamination before further processing for microbial analyses, all surfaces of the frozen sections were removed before nucleic-acid extraction. The permafrost soil cubes were sprayed with 1.82×10^9 fluorescent microbeads (Fluoresbrite Yellow Green latex microbeads, 0.5 μm, Polyscience) to detect removal of the possibly contaminated surface soil of the cubes. The beads were washed three times with 70% RNase-free ethanol and twice with diethylpyrocarbonate (DEPC)-treated Milli-Q water before use. All tools were sterilized by autoclaving twice followed by baking at 120 °C overnight. The cubes were sawn with an electronic jigsaw on a dry ice bed using sterile blades. Sterile, chilled clamps were used to hold the cube as it was sawn. After cutting a side of the cube, the clamps were soaked in 10% bleach followed by sterile Milli-Q water and finally wiped with 70% ethanol before using them for sawing the new surface. During the sawing, the dry ice was covered with sterile and RNase-free foil that was changed after sawing each of the cube's sides. After sawing, samples were taken from the newly exposed surface with a sterile scalpel and examined by epifluorescence microscopy for the presence of the fluorescent beads. Acridine orange direct staining and epifluorecence microscopy was conducted on all samples; cell density in permafrost was found to be approximately 1×10^8, and in bog and active layer was twice that amount: approximately 2×10^8 cells per gram of soil (data not shown).

Subsamples were taken from the cleaned and frozen soil cubes, homogenized while frozen and nucleic acids were extracted from three replicate 500 mg of soil from each cube. Nucleic acids were extracted using the hexadecyltrimethyl ammonium bromide (CTAB) and phenol–chloroform and bead beating protocol (see ref. 3) with a modification of adding a pressure lysis step before bead beating with a Barocycler (Pressure Biosciences) set to alter pressure from ambient to 2,413 bar for 20 s and back to ambient for 10 s, 20 times. After extraction, nucleic acids were purified with the AllPrep kit (Qiagen) according to the manufacturer's instructions, with DNase treatment for the RNA samples. The purified replicates of the DNA and RNA samples were pooled and the resulting yield and quality were measured with a Bioanalyzer (Agilent).

The V6–V9 area of the 16S rRNA gene was amplified at the Joint Genome Institute (JGI) with primers 926wF (aaa ctY aaa Kga att gRc gg) and 1406R creating an amplicon of approximately 480 base pairs (bp). The reverse primer included a 5-bp barcode for multiplexing of samples during sequencing. The sequence reads were de-noised using Pyrotagger[19]. Sequence reads were assigned into OTUs with 97% similarity. The representative from each OTU (97%) was classified by comparing with the Greengenes database (version 29nov2010). Potential chimaeras were removed using Pyrotagger and 69 OTUs were detected as possibly chimaeric and removed from the data set.

As the number of sequence reads recovered from each sample varied, the sample data sets were rarified before alpha diversity analysis using the number of reads in the lowest sample. The microbial diversity did not vary drastically between the different samples, showing the frozen and thawed sites to carry a diverse microbial community by the OTU number and chao1 and ACE diversity estimates (Extended Data Table 2).

Alpha diversity indices were calculated with R statistical calculation software package vegan (http://cran.r-project.org/web/packages/vegan/index.html)[20] and in QIIME[21] using a rarified data set as the number of sequence reads per sample varied.

The OTU matrix with the relative abundance of each OTU was used to build a rank abundance curve of the 20 most abundant OTUs in all samples (Extended Data Fig. 5). The 454 reads have been deposited in the Sequence Read Archive under BioProject number PRJNA222786.

Metagenomes. Two samples from each of the three sample types were metagenome sequenced: active layer samples A2 and A5, permafrost samples P1 and P3 and thermokarst bog samples B3 and B4. The quantity of the extracted DNA from the permafrost samples was not adequate for Illumina sample preparation and thus the DNA was emulsion PCR amplified as previously described[3] after shearing it into fragments of about 300–500 bp. Briefly, 10 ng of DNA was sheared with a Covaris shearing apparatus and Illumina sequencing adapters were ligated to the end-repaired fragments. The fragments were amplified with the adaptor primers in an emulsion so that each fragment was amplified in an individual oil droplet. After amplification, the correct size fragments were extracted from a low-melting electrophoresis gel; the purified fragments were sequenced using the Illumina GAII sequencing platform (Illumina) for two cycles, generating 2 × 113 bp paired-end reads.

Assembly. The metagenomic reads were assembled using the *de novo* assembler at CLC Genomics Workbench (CLC bio). The default parameters were used except the minimum length of contigs, which was 1,000. CLC assembler resulted in assembly of 18.8–59.8% of reads with an N50 from 890 to 1,117 bp.

All of the contigs and the individual reads were submitted to the Integrated Microbial Genomes and Metagenomes (IMG/M, US Department of Energy[22]) for gene calling and annotation. Additionally, the reads were annotated in MG-RAST[23] against the SEED subsystems database with an e value of 10^{-5}.

Binning draft genomes from MG. Reads were mapped to contigs using BWA. Contigs greater than 2.5 kilobases in length were grouped into genome bins on the basis of abundance and tetranucleotide frequency using MetaBat (https://bitbucket.org/berkeleylab/metabat). To account for the fact that our data represent complex communities and to minimize contamination, we used the very specific option, which specifies a probability cutoff of 90% for imminent neighbours, 90% for secondary neighbours and 90% for tertiary neighbours. All other parameters were set to default. Completeness and contamination of genome bins were assessed using

CheckM[24], which identified lineage-specific single-copy marker genes in each bin. Genome completeness was estimated using number of marker genes present in genome and contamination was evaluated from the number of multi-copy marker genes. Bins that were less than 20% complete or with greater than 10% contamination were discarded. AMPHORA2 was used to evaluate taxonomy. A confidence score for each gene was generated at each rank of classification[25]. Marker lineage was reported if 75% of the classifications were in agreement at a particular taxonomic level. Genome-to-Genome Distance Calculator[11] was used for selected binds to calculate the DDH similarity. The annotated bins are stored at IMG/M under Study ID Gs0063124.

Metatranscriptomics. Approximately 14 ng of RNA was used in first-strand synthesis of complementary DNA (cDNA) with 100 nM of primer N6-T7 (5′ NNNN NN-T7) and Superscript III First-Strand kit (Invitrogen) at 37 °C without the last denaturation step[26] and with 200 μM of random hexamers. An Invitrogen Second-Strand synthesis kit was used for second-strand synthesis at 16 °C followed by purification with Qiagen PCR purification columns. The purified cDNA was amplified with the Ambion aRNA amplification kit (Ambion) according to the manufacturer's protocol. The resulting cDNA was purified on Qiagen PCR columns and eluted to RNase free water. The cDNA was quantified and qualified with Bioanalyzer Pico RNA chips (Bioanalyzer).

For the permafrost samples, 10 ng of linear amplified cDNA was used for Illumina library preparation through emulsion PCR as with metagenomes. The active layer and thermokarst samples were used directly in the Illumina library preparation at the JGI sequencing facility using the manufacturer's protocols. One Illumina HiSeq lane was sequenced per sample with 150 bp for two cycles generating 2 × 150 bp paired-end reads. Individual reads were annotated at IMG/M and MG-RAST according to the parameters used for metagenomic reads. Metagenomic and metatranscriptomic read annotations can be viewed at IMG/M under following submission numbers: sample P1-MG, 10301 and 10302; P3-MG, 10298 and 10299; B3-MG, 10560; B4-MG, 10561 and 10562; A2-MG, 10308 and 10559; A5-MG, 10305 and 10306; P3-MT, 10300; B3-MT, 10304; B4-MT, 10303; A2-MT, 10309; A5-MT, 10307.

Metagenomic and metatranscriptomic annotations can be viewed at IMG/M under Study ID Gs0063124.

Metaproteomics. Proteins were extracted using the SDS lysis-based method with trichloroacetic acid (TCA) precipitation and cleanup step followed by 10 kDA spin filter, modified from ref. 9. Briefly, 5 g of frozen sample were suspended in 5 ml of SDS lysis buffer (4% SDS, 100 mM Tris-HCl, pH 8.0) and boiled for 5 min; this procedure presumably shuts down all cellular processes and captures the metaproteome *in situ*. Cells were further disrupted by sonication pulses of 10 s on, 10 off, for 2 min. After the second 5 min boiling step the samples were centrifuged and dithiothreitol (DTT) was added to a final concentration of 24 mM to the supernatant. The samples were precipitated with 20% TCA overnight at −20 °C. Precipitated samples were centrifuged at 10,000*g* for 40 min at 4 °C and the pellets were washed with ice-cold acetone. After mixing by vortexing, the samples were centrifuged for 5 min at 4 °C at 21,000*g*. Protein pellets were solubilized in 8 M urea with 100 mM Tris-HCl, pH 8.0, by incubation at room temperature (27 °C) for 30 min and sonicated in an ice-water bath to keep the samples below 37 °C. Samples were adjusted to 20 mM iodoacetamide and incubated in the dark for 15 min at room temperature. Trypsin digestion and peptide clean-up were performed as in ref. 27; the resultant peptide solutions were split into four replicates and frozen at −80 °C until analysis.

Measurement of peptides by two-dimensional liquid chromatography–tandem mass spectrometry. Complex peptide samples from each soil type were pressure cell loaded onto a biphasic two-dimensional back column packed with 5 cm strong cation exchange resin for charge-based separation of peptides followed by 3 cm C18 reversed phase for online desalting (Luna and Aqua respectively, Phenomenex). Once loaded, the sample columns were washed offline with solvent A (5% acetonitrile, 95% high-performance liquid chromatography (HPLC)-grade water, 0.1% formic acid) for 15 min to remove residual urea and NaCl followed by a gradient to 100% solvent B (70% acetonitrile, 30% HPLC-grade water, 0.1% formic acid) over 30 min to move the peptide population from the reversed phase to the strong cation exchange resin. Washed samples were then placed in-line with a nanospray emitter (100 μm internal diameter, New Objective) packed with 15 cm of C18 reversed phase material (Aqua) and analysed in duplicate by 24 h two-dimensional liquid chromatography–tandem mass spectrometry (MS/MS) (11 salt-pulses: 5, 7.5, 10, 12.5, 15, 17.5, 20, 25, 35, 50, 100% of 500 mM ammonium acetate followed by a 100 min gradient to 50% solvent B) with a hybrid Velos/Orbitrap mass spectrometer (Thermo Fisher Scientific) operating in data-dependent mode. Full MS1 scans were obtained using an Orbitrap mass analyser set to 330,000 resolution, while 20 data-dependent MS/MS scans were acquired per full scan (two microscans). Data-dependent MS/MS scans were acquired in the Velos linear ion trap, with two microscans averaged and dynamic exclusion set to on. Two replicate measurements were obtained for each sample, except sample B4 which had one measurement.

Database searching and protein identification. FASTA databases were compiled with the protein sequences of the assembled and annotated metagenomes (active layer, permafrost and thermokarst bog), 180 proteomes of known soil bacteria (Supplementary Table 2) and known protein contaminants (trypsin, keratin, etc.) into one single composite database. Assembled metagenomes used for proteome searches were submitted to IMG under the following submission numbers: permafrost, 2321; active layer, 2382; thermokarst bog, 2383.

All MS/MS spectra were searched as in ref. 27 with the SEQUEST algorithm against the composite FASTA database and filtered with DTASelect/Contrast at the peptide level (Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)). Only proteins identified with two fully tryptic peptides, with one unique to that specific protein entry, from a 22 h run were considered for further biological study.

Metaproteomics data analysis. To prepare for semiquantitative proteome analysis, DTASelect-filtered data were subjected to spectral count balancing and subsequent normalized spectral abundance factor (NSAF) determination[28]. For the former, the unique status of each identified peptide was assessed. If a peptide was deemed unique, it retained 100% of its previously assessed spectral count. However, if a peptide was found to belong to more than one representative protein, namely non-unique, its spectral count was re-calculated on the basis of the ratio of uniquely identified peptides between the two or more proteins that shared the non-unique peptide in question. The adjusted spectral count of the proteins incorporated these balanced values, which corrected for the slight quantitative bias that occurs with homologous proteins and/or proteins with homologous regions, both of which artificially inflate spectral count values at the expense of proteins that share no homology. Once spectrally balanced, NSAF values were calculated for each protein in a specific run to normalize individual MS runs on the basis of the total number of spectra collected and protein length, which by itself introduces a bias that favours larger proteins if not corrected for. Summed technical replicates and NSAF values were used for subsequent analysis. A Venn diagram was drawn for the shared and unique proteins present in three out of four technical replicates in each soil. The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD001131.

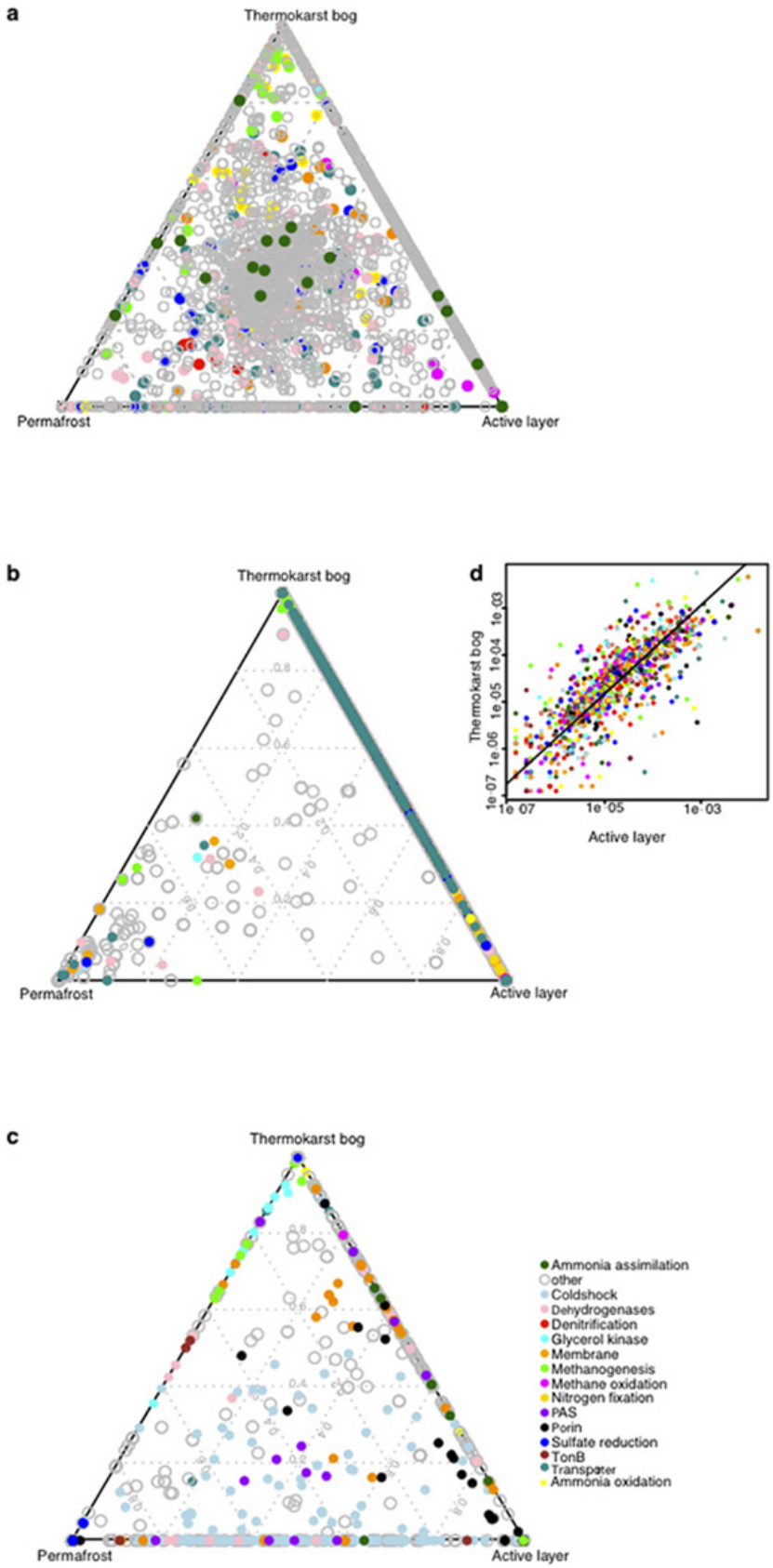
Proteome search to additional genomes. Microbial proteins from 13 recently sequenced species isolated from cold environments (Supplementary Table 2) were collected from NCBI RefSeq and JGI in FASTA format. With these protein sequences, we searched MS/MS peaks of our data using MyriMatch 2.1.138. We set MyriMatch options as MonoPrecursorMzTolerance = 3 *m*/*z*, MonoisotopeAdjustmentSet = 0, FragmentMzTolerance = 0.8 *m*/*z*, CleavageRules = Trypsin/P, MinTerminiCleavages = 1, ComputeXCorr = true and MaxResultRank = 10. We also added static carbamidomethyl modification on Cys by adding 57.021462 Da. We ran MyriMatch with 48 central processing units at the Newton cluster of the University of Tennessee, Knoxville (https://newton.utk.edu/). PepXML files, produced by MyriMatch, were analysed by in-house Python scripts to get proteins from peptide-to-spectrum matches as was done by DTASelect in the original search. Spectrum counts of every protein were balanced and normalized by NSAF as was done in the original search. BLASTP was used to compare protein sequences between the original database and 13 new species. We used default options of BLASTP and accepted BLAST hits that their matched length was bigger than 95% length of a query protein and had no gap opens.

Data analysis. In addition to IMG/M annotation, a BLAST search was conducted on the metagenomic and metatranscriptomic reads against the Greengenes database (version 29nov2010); one best hit was chosen for each read, with an *e* value cutoff of 10^{-5} to focus on the 16S rRNA gene profiles. Since the Illumina HiSeq runs produced over 20 Gb of data, the reads were pre-analysed with the Ribopicker program to collect only the presumably 16S rRNA origin reads, which were further classified with BLAST. The relative abundance of each phylum in MG and MT data sets was compared in STAMP[29] with Fisher's alpha test and Storey's false discovery rate correlation. MT/MG data were compared with a two-tailed *t*-test.

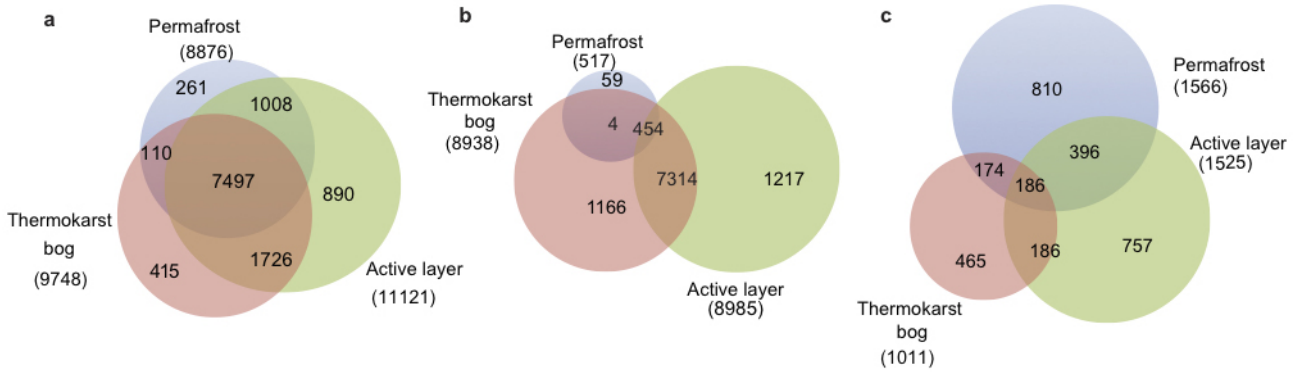
Triplots based on the relative abundance, normalized read annotations were calculated in R package vcd (http://cran.r-project.org/web/packages/vcd/index.html)[30] and barcharts were drawn for selected cycles using the normalized data. The heatmap focusing on the transporters was obtained from the top 40 (relative abundance) MT reads that were assigned as transporters. A Venn diagram (Extended Data Fig. 2) was constructed from the selected genes present in the three sample types (permafrost, active layer and bog) for all three data sets (MG, MT and MP).

The three different omics approaches were compared (Extended Data Fig. 6) for each site by extracting the Kyoto Encyclopedia of Genes and Genomes (KEGG) hits for MG, MT and MT data sets and comparing these in ternary plots as described above; the MT and MG data sets were multiplied by 100 to make the values comparable.

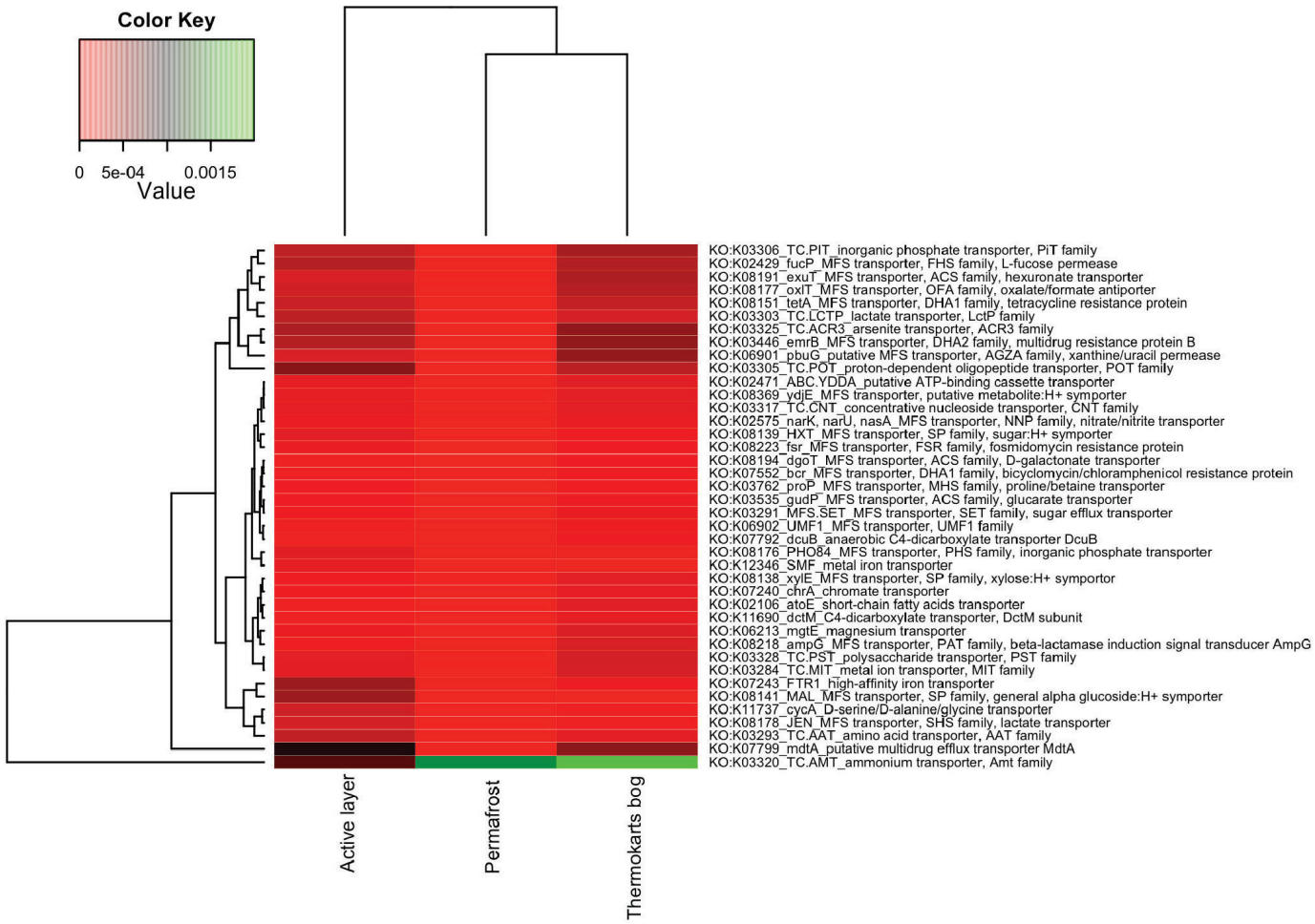
14. Lovley, D. R. & Phillips, E. J. Rapid assay for microbially reducible ferric iron in aquatic sediments. *Appl. Environ. Microbiol.* **53,** 1536–1540 (1987).
15. Balderston, W. L., Sherr, B. & Payne, W. J. Blockage by acetylene of nitrous oxide reduction in *Pseudomonas perfectomarinus*. *Appl. Environ. Microbiol.* **31,** 504–508 (1976).
16. Balch, W. E., Fox, G. E., Magrum, L. J., Woese, C. R. & Wolfe, R. S. Methanogens: reevaluation of a unique biological group. *Microbiol. Rev.* **43,** 260–296 (1979).
17. Zehnder, A. J. & Brock, T. D. Anaerobic methane oxidation: occurrence and ecology. *Appl. Environ. Microbiol.* **39,** 194–204 (1980).
18. Macy, J. M., Snellen, J. E. & Hungate, R. E. Use of syringe methods for anaerobiosis. *Am. J. Clin. Nutr.* **25,** 1318–1323 (1972).
19. Kunin, V. & Hugenholtz, P. PyroTagger: a fast, accurate pipelinde for analysis for analysis of rRNA amplicon pyrosequencing data. *Open J.* **1,** 1–8 (2010).
20. Oksanen, J., Kindt, R., Legendre, P. & O'Hara, R. B. Vegan: Community Ecology Package v.1.8-2 (2006).
21. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
22. Markowitz, V. M. *et al.* IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40,** D115–D122 (2012).
23. Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
24. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *PeerJ PrePrints* **2,** e554v1 (2014).
25. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28,** 1033–1034 (2012).
26. Leininger, S. *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442,** 806–809 (2006).
27. Giannone, R. J. *et al.* Proteomic characterization of cellular and molecular processes that enable the *Nanoarchaeum equitans–Ignicoccus hospitalis* relationship. *PLoS ONE* **6,** e22942 (2011).
28. Florens, L. & Washburn, M. P. Proteomic analysis by multidimensional protein identification technology. *Methods Mol. Biol.* **328,** 159–175 (2006).
29. Parks, D. H. & Beiko, R. G. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26,** 715–721 (2010).
30. Meyer, D., Zeileis, A. & Hornik, K. vcd: Visualizing Categorical Data v.1.2-13 (2012).



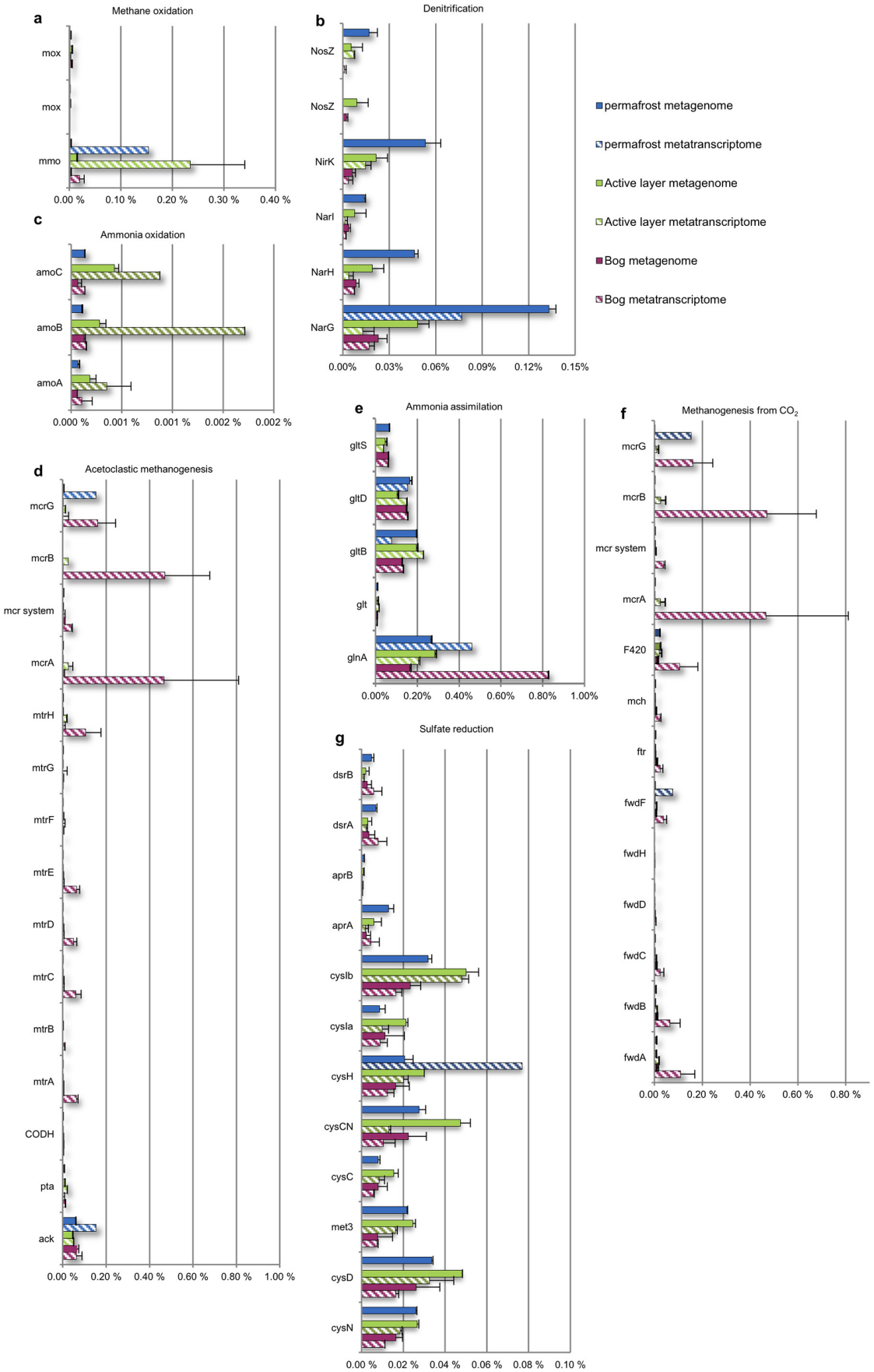
Extended Data Figure 1 | Comparison of multi-omics data between the three zones. The triplots visually present selected functions in (**a**) metagenomes, (**b**) metatranscriptomes and (**c**) metaproteomes in the three studied zones (permafrost, active layer and thermokarst bog). The colours correspond to SEED subsystems categories in MG-RAST; in addition, the closer the symbol is to the node of the triplot, the more abundant the gene, transcript or protein is in that particular soil type. Functions in the centre are shared among all three sites, those on edges are shared between two zones and those on the nodes are unique to that site. Owing to the large number of shared functions between bog and active layer in the metatranscriptomes, these categories are specifically compared in **d**.



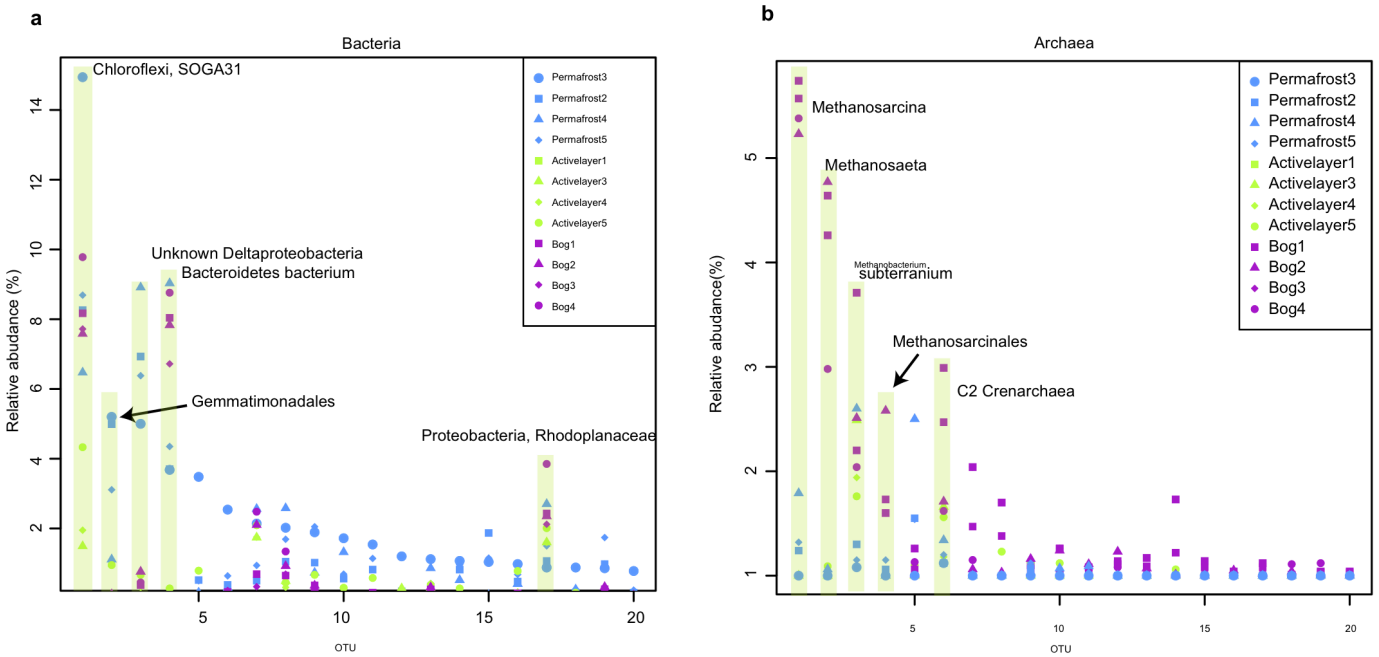
Extended Data Figure 2 | Shared and unique genes, transcripts and proteins in the three zones. The Venn diagrams for (**a**) metagenome, (**b**) metatranscriptome and (**c**) metaproteome data illustrate the shared and unique genes, transcripts and proteins in the three zones.



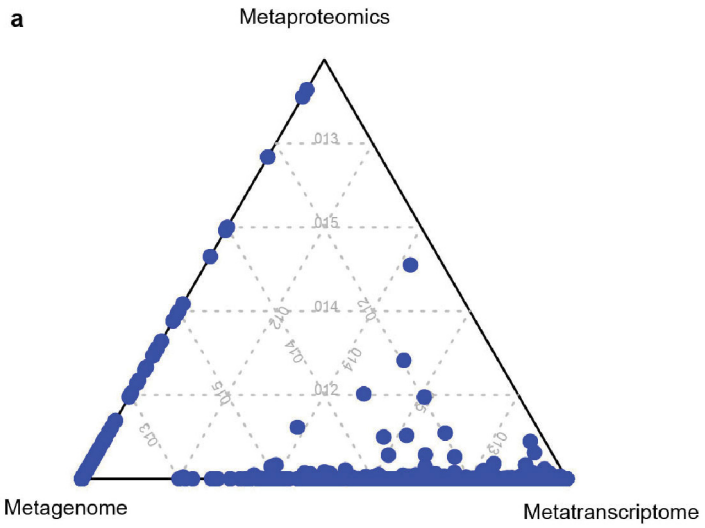
Extended Data Figure 3 | Heatmap showing the 40 most abundant transporters found in the metatranscriptomes.

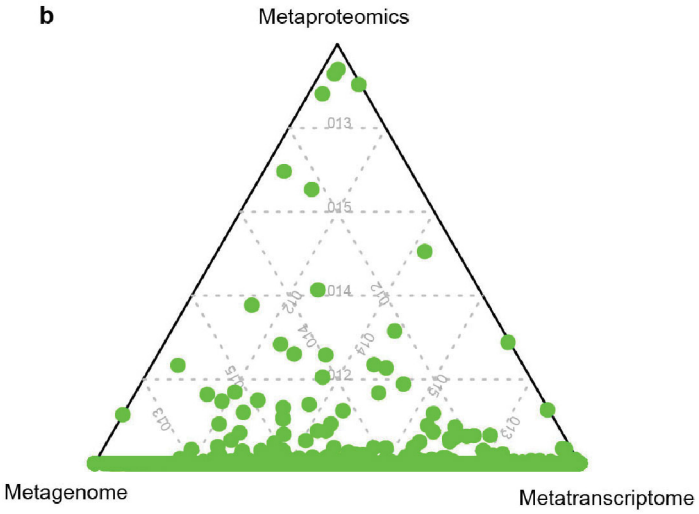


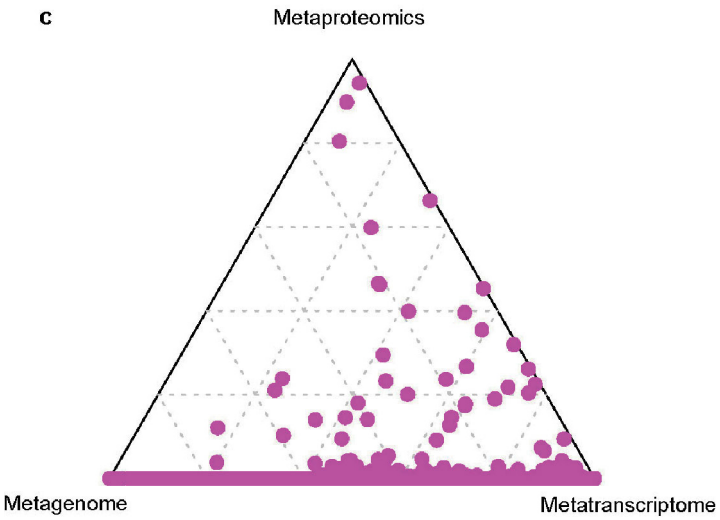
Extended Data Figure 4 | Abundance of genes involved in sulphate reduction, nitrogen cycle, methanogenesis and methane oxidation. The relative abundance of each gene is shown; error bars, s.d. The genes with protein matches in the metaproteome data are marked with a triangle.



Extended Data Figure 5 | The 20 most abundant OTUs found in permafrost, active layer and thermokarst bog. **a**, Bacterial OTUs; **b**, archaeal OTUs. A single Chloroflexi OTU constituted up to 15% of the OTUs in permafrost and a draft genome bin was obtained that corresponded to this OTU (Extended Data Table 1). The abundant archaeal OTUs were methanogens.







Extended Data Figure 6 | Comparison of abundant KEGG orthologous groups found in all three omics approaches (MG, MT and MP) for each site. Permafrost: restriction modification systems, PAS sensors, ABC transporters for oligopeptides and nucleic-acid synthesis. Active layer: chaperonins, dehydrogenases and transporters for branched-chain amino acids and sugars. Thermokarst bog: methyl coenzyme M reductase, glycerol kinase, nucleic-acid synthesis and ATPases.

Extended Data Table 1 | Metagenomic bins produced from the metagenomes

			Marker Gene Copies								Genes of interest																								
													methanogenesis							transporter for																
Zone	Bin ID	Marker Lineage	0	1	2	3	Completeness	Contamination	Size	Contig Count	cysN/C	dsrA-B	cytochrome b/c	fmdA-E	mtrA, C-E, G	mcrA, B, G	phosphotransacetylase	CO dehydrogenase/acetyl-CoA synthase	methyl-H4SPT:coenzyme M methyltransferase	Fe3+	S	Mg/Co/Ni	Mn	P	aminoacids	NO3/NO2	NH4	Xylose	Thiamin	Molybdenum	oligopeptide/dipeptide	similar DDH %
Active Layer	AL_334	p_Proteobacteria, c_Betaproteobacteria	85	19	0	0	24.83	0	1,497,758	371			x							x				x	x					x	x	
Permafrost	PF_55	p_Gemmatimonadetes, c_Gemmatimonadetes, o_Gemmatimonadales, f_Gemmatimonadaceae, g_Gemmatimonas, s_Aurantiaca	82	21	1	0	31.9	1.72	522,584	134																				x	x	
	PF_400	p_Actinobacteria, c_Actinobacteria, o_Acidimicrobiales, f_Acidimicrobiaceae, g_Acidimicrobium, s_Ferrooxidans	17	152	2	0	89.74	1.28	4,300,461	237	x		x							x	x	x	x	x	x	x	x	x	x	x	x	*Acidimicrobium ferrooxidans*, DDH 19.60% +/- 2.30
	PF_8	p_Proteobacteria	136	55	0	0	27.67	0	454,972	118															x							
Bog	Bog_15	p_Euryarchaeota, c_Methanomicrobia	23	201	5	0	88.15	2.92	1,227,542	154				x	x	x		x		x			x	x	x		x					*Methanosaeta concilii* DDH 42.70 +/- 2.53
	Bog_232	p_Euryarchaeota, c_Methanomicrobia	44	183	2	0	77.14	0.97	879,929	153				x				x						x	x		x			x		*Methanosaeta concilii*, DDH 46.60% +/- 2.58
	Bog_301	p_Euryarchaeota, c_Methanomicrobia	62	167	0	0	70.47	0	1,008,045	190				x	x					x		x		x	x		x			x	x	*Methanosarcina acetivorans*, DDH 69.70% +/- 2.92
	Bog_263	p_Chloroflexi	45	96	10	0	62.42	8.42	2,081,309	490	x	x	x							x				x	x					x	x	
	Bog_1285	p_Spirochaetes, c_Spirochaetia, o_Spirochaetales, f_Spirochaetaceae	65	75	1	1	54.19	2.38	2,013,476	473		x	x							x				x	x		x	x		x	x	
	Bog_440	p_Chloroflexi, c_Anaerolineae, o_Anaerolineales, f_Anaerolineaceae, g_Anaerolinea, s_Thermophila	70	76	17	0	47.95	9.12	1,269,542	328		x											x	x						x	x	*Anaerolinea thermophila*, DDH 15.20% +/- 2.14
	Bog_246	p_Proteobacteria	75	28	1	0	30.25	1.72	1,127,004	295												x			x							
	Bog_283	p_Proteobacteria, c_Deltaproteobacteria, o_Myxococcales	172	64	12	0	28.36	4.56	1,845,531	529	x		x							x					x		x	x			x	
	Bog_474	p_Actinobacteria, c_Actinobacteria, o_Solirubrobacterales, f_Conexibacteraceae, g_Conexibacter, s_Woesei	83	21	0	0	23.24	0	886,523	220														x	x							
	Bog_532	p_Proteobacteria, c_Deltaproteobacteria, o_Syntrophobacterales, f_Syntrophaceae, g_Syntrophus, s_Aciditrophicus	188	58	2	0	23.08	0.36	540,522	151															x							

x gene present
x RNA transcripts
x proteins

cysN/C: sulfate adenylyltransferase
dsr: dissimilatory sulfate reduction
fmd: formylmethanofuran dehydrogenase
mtr: Tetrahydromethanopterin S-methyltransferase
N5-methyltetrahydromethanopterin:coenzyme M methyltransferase
mcr: Methyl coenzyme M reductase

Extended Data Table 2 | Number of OTUs from 454 sequencing and alpha diversity indices

Sample	sequences after QC	ACE	chao1 lower bound	chao1 upper bound	dominance	OTUs	singles	doubles	reciprocal simpson	shannon	simpson	simpson eveness
Active_layer_A2	13196	3226.70	2798.70	3397.00	0.008	1544	827	223	1.008	8.306	0.992	0.001
Active_layer_A3	2790	1076.23	955.00	1368.01	0.016	524	289	68	1.016	7.322	0.984	0.002
Active_layer_A4	9166	1939.81	1687.22	2077.29	0.014	1048	521	166	1.015	7.630	0.986	0.001
Active_layer_A5	9217	3436.39	2831.32	3571.33	0.009	1361	815	183	1.010	8.128	0.991	0.001
Bog_site_B1	10430	2332.45	2071.78	2532.03	0.026	1234	630	189	1.026	7.492	0.974	0.001
Bog_site_B2	9455	1800.56	1621.60	2010.99	0.024	1015	485	150	1.025	7.332	0.976	0.001
Bog_site_B3	8317	2039.81	1745.67	2127.27	0.024	1097	546	181	1.024	7.522	0.976	0.001
Bog_site_B4	8194	1611.11	1369.57	1726.94	0.031	853	428	135	1.032	6.913	0.969	0.001
Permafrost_layer_P1	7888	5228.38	4100.70	5016.39	0.019	1810	1208	268	1.020	8.195	0.981	0.001
Permafrost_layer_P3	20442	2333.84	2280.85	2901.55	0.035	1242	609	140	1.036	6.959	0.965	0.001
Permafrost_layer_P4	8935	1846.10	1575.33	2051.06	0.030	875	473	122	1.031	6.826	0.970	0.001
Permafrost_layer_P5	24306	2438.27	2220.88	2720.56	0.021	1333	622	173	1.022	7.274	0.979	0.001

LETTER

doi:10.1038/nature14243

Pathogen-secreted proteases activate a novel plant immune pathway

Zhenyu Cheng[1,2]*, Jian-Feng Li[1,2]*†, Yajie Niu[1,2], Xue-Cheng Zhang[1,2], Owen Z. Woody[3], Yan Xiong[1,2]†, Slavica Djonović[1,2]†, Yves Millet[1,2]†, Jenifer Bush[1], Brendan J. McConkey[3], Jen Sheen[1,2] & Frederick M. Ausubel[1,2]

Mitogen-activated protein kinase (MAPK) cascades play central roles in innate immune signalling networks in plants and animals[1,2]. In plants, however, the molecular mechanisms of how signal perception is transduced to MAPK activation remain elusive[1]. Here we report that pathogen-secreted proteases activate a previously unknown signalling pathway in *Arabidopsis thaliana* involving the Gα, Gβ, and Gγ subunits of heterotrimeric G-protein complexes, which function upstream of an MAPK cascade. In this pathway, receptor for activated C kinase 1 (RACK1) functions as a novel scaffold that binds to the Gβ subunit as well as to all three tiers of the MAPK cascade, thereby linking upstream G-protein signalling to downstream activation of an MAPK cascade. The protease–G-protein–RACK1–MAPK cascade modules identified in these studies are distinct from previously described plant immune signalling pathways such as that elicited by bacterial flagellin, in which G proteins function downstream of or in parallel to an MAPK cascade without the involvement of the RACK1 scaffolding protein. The discovery of the new protease-mediated immune signalling pathway described here was facilitated by the use of the broad host range, opportunistic bacterial pathogen *Pseudomonas aeruginosa*. The ability of *P. aeruginosa* to infect both plants and animals makes it an excellent model to identify novel immunoregulatory strategies that account for its niche adaptation to diverse host tissues and immune systems.

We found that culture filtrate of *P. aeruginosa* strain PA14 activates an *Arabidopsis* β-glucuronidase (GUS) reporter gene under the control of the pathogen-inducible *CYP71A12* promoter (*CYP71A12pro:GUS*). Whereas the well-characterized immune elicitor flg22, a synthetic peptide that corresponds to the active epitope of bacterial flagellin, induces *CYP71A12pro:GUS* in the root elongation zone[3], PA14 culture filtrate activates the reporter in the cotyledons and leaves of both wild-type *Arabidopsis* Col-0 and *fls2* mutant seedlings in which the flagellin receptor is mutated (Fig. 1a).

By screening a collection of 64 *P. aeruginosa* PA14 regulatory and secretion-related mutants, we found that the induction of the *CYP71A12* promoter was dependent on the quorum-sensing gene *lasI* and on the type II secretion apparatus-encoding genes *xcpR*, *xcpT*, *xcpW*, and *xcpZ* (Fig. 1a and Extended Data Table 1). Ion-exchange chromatography fractionation (Extended Data Fig. 1a) followed by mass spectrometry (data not shown) identified the elicitor in the PA14 secretome as protease IV, a type II-secreted, PvdS-regulated lysyl class serine protease

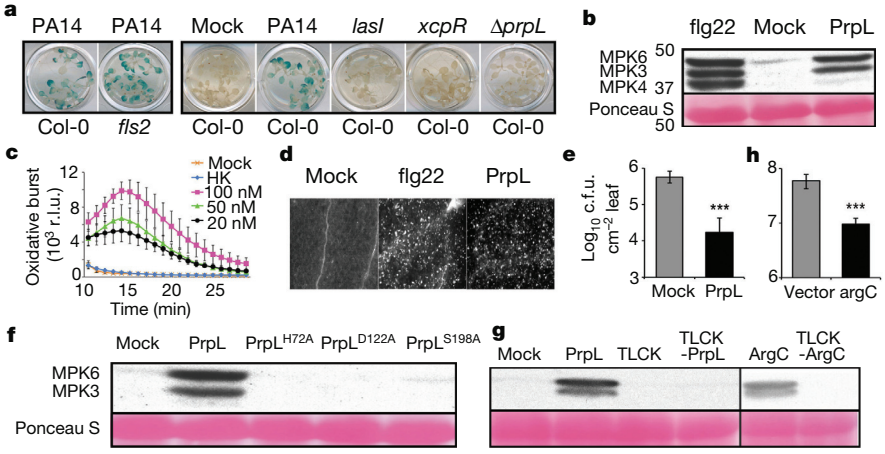


Figure 1 | Proteases trigger innate immune responses in *Arabidopsis* via proteolytic activity. **a**, Activation of *CYP71A12pro:GUS* in wild-type *Arabidopsis* Col-0 or *fls2* mutant cotyledons by culture filtrates from wild-type *P. aeruginosa* PA14, from PA14 mutants containing a transposon insertion in *lasI* or *xcpR*, or from PA14/Δ*prpL*. **b**, Western blot depicting activation of MAPKs by PrpL or flg22. Numbers on the left axis of the blot represent marker size (molecular mass in kilodaltons). **c**, Chemiluminescence assay showing elicitation of an oxidative burst by PrpL; r.l.u, relative luminescence units. HK: 100 nM 'heat-killed' PrpL. **d**, Callose formation in cotyledons elicited by PrpL or flg22 detected by aniline blue staining. **e**, Protection of 4-week-old *Arabidopsis* leaves from *P. syringae* pv. *tomato* strain DC3000 infection by pre-infiltrated PrpL; c.f.u., colony-forming units. **f**, Western blot depicting activation of MPK3 and MPK6 by PrpL and inactive variants of PrpL. The same molecular mass region of the blot is shown as in **b**. **g**, Western blot depicting activation of MPK3 and MPK6 by PrpL or ArgC or TLCK-treated PrpL or TLCK-treated ArgC. The same molecular mass region of the blot is shown as in **b**. **h**, Growth of *X. campestris* strains 8004/*argC* or 8004/vector in 3-week-old *B. oleracea* leaves. Data represent mean ± s.d.; $n = 16$ individual seedlings (**c**) and $n = 10$ leaves from five plants (**e**, **h**); $^{***}P < 0.001$, Student's *t*-test. The experiments in **a** and **d** were repeated three times with similar results and the representative images shown were selected from at least three images.

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Department of Biology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. †Present addresses: State Key Laboratory of Biocontrol and Guangdong Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275, China (J.-F.L.); Shanghai Center for Plant Stress Biology, Chinese Academy of Sciences, Shanghai, 201602, China (Y.X.); Symbiota, Inc., 100 Edwin Land Boulevard, Cambridge, Massachusetts 02142, USA (S.D.); Synlogic, 130 Brookline Street, Cambridge, Massachusetts 02139, USA (Y.M.).
*These authors contributed equally to this work.

encoded by the *P. aeruginosa prpL* gene (PA14_09900). Purified His-tagged PA14 protease IV (referred to as PrpL in the figure legends) activated *CYP71A12pro:GUS* (Extended Data Fig. 1b), whereas culture filtrate from an in-frame deletion mutant of *prpL* (PA14/Δ*prpL*) did not (Fig. 1a).

Purified protease IV is a very strong elicitor of immune responses in *Arabidopsis*, comparable to flg22 in the activation of MPK3 and MPK6 (but not MPK4) (Fig. 1b), elicitation of an oxidative burst (Fig. 1c), deposition of callose in cotyledons (Fig. 1d), and protection of adult *Arabidopsis* leaves from *Pseudomonas syringae* pathovar (pv.) *tomato* strain DC3000 infection (Fig. 1e). In contrast, trypsin, a well-characterized serine protease, failed to activate MAPK cascades or trigger an oxidative burst (Extended Data Fig. 2a, b). Global transcriptional profiling analysis (Extended Data Fig. 3a), confirmed by quantitative PCR with reverse transcription (RT-qPCR) analysis of selected defence-related genes (Extended Data Fig. 3b), showed a high degree of concordance between the genes activated or repressed by protease IV and genes previously shown to be regulated by flg22 or oligogalacturonides in seedlings[4] (Pearson correlation coefficients of 0.899 and 0.864 for protease-IV-treated versus flg22 and oligogalacturonides, respectively).

Importantly, protease IV variants containing alanine substitutions at the proteolytic catalytic triad site ($PrpL^{H72A}$, $PrpL^{D122A}$, $PrpL^{S198A}$), which exhibit no detectable proteolytic activity[5], were impaired for MAPK activation (Fig. 1f), defence gene induction, and oxidative burst elicitation (Extended Data Fig. 4a, d). Treatment of protease IV with the protease inhibitor TLCK (Fig. 1g and Extended Data Fig. 4b, d) or with heat (Fig. 1c and Extended Data Fig. 4c) also resulted in a loss of elicitation ability.

The closest homologue of *P. aeruginosa* protease IV in sequenced bacterial genomes is encoded by the *argC* gene of *Xanthomonas campestris*, a bona fide plant pathogen (Extended Data Fig. 5a). Purified His-tagged ArgC protease exhibited protease activity *in vitro* and triggered the activation of MPK3 and MPK6 that is dependent on ArgC protease activity (Fig. 1g).

We noticed that there is a high rate of naturally occurring null mutations in the *Xanthomonas argC* gene (8 out of 22 total alleles in sequenced *Xanthomonas* genomes; Extended Data Fig. 5b–d), suggesting that *argC* is probably under negative selection. Consistent with the sequence data, the culture filtrate of strain *X. campestris* pv. *raphani* strain 1946, from which the functional *argC* gene was cloned, activated the *CYP71A12pro:GUS* reporter, whereas culture filtrates from two *X. campestris* pv. *campestris* strains (8004 and BP109), which contain presumptive *argC* null frame shift mutations, failed to activate (Extended Data Fig. 5e). We complemented the null *argC* mutant in strain 8004 (*Xcc* 8004) with the functional *argC* gene from strain 1946 (8004/*argC*) (Extended Data Fig. 5e). Consistent with ArgC-mediated induction of a host immune response during an infection in a mature plant, the growth of 8004/*argC* in *Brassica oleracea* (broccoli), a natural host of *X. campestris*, was reduced about sixfold compared with the 8004/vector control (Fig. 1h). The expression of haemagglutinin (HA)-tagged ArgC was readily detected in broccoli leaves infected with 8004/*argC* (Extended Data Fig. 5e), indicating that ArgC is synthesized during infection.

Next, we sought to investigate the mechanism by which protease IV activates an immune response in *Arabidopsis*. Previous studies have shown that G proteins play a role in microbe-associated molecular pattern molecule-mediated responses[6]. In the case of protease IV, we found reduced expression of defence-related genes in *gα* or *gβ* mutants (and in a *gγ1gγ2* double mutant), reduced levels of the oxidative burst in a *gα* mutant and a *gαβ* double mutant, reduced MPK3 and MPK6 activation, and reduced protection against *P. syringae* infection in a *gαβ* double mutant (Fig. 2a–c and Extended Data Fig. 6a, b). The induction of *CYP71A12* and activation of MPK3 and MPK6 by *X. campestris* ArgC was also diminished in the G-protein mutants, similar to the pattern observed for protease IV (Fig. 2a, b). In contrast to protease IV and ArgC, in the case of flg22, defence gene expression was only reduced in *gβ* and *gαβ* double mutants, the oxidative burst was more severely

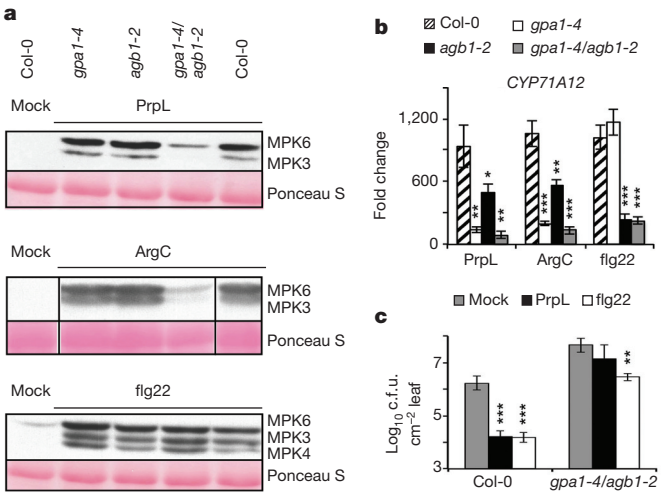


Figure 2 | Protease-mediated defence responses are coupled to G-protein signalling. **a**, Western blot depicting activation of MAPKs by PrpL, ArgC, or flg22 in wild-type Col-0 or G-protein tDNA mutants. The same molecular mass region of the blot is shown as in Fig. 1b. **b**, Induction of defence-related gene expression by PrpL, ArgC, or flg22 in wild-type Col-0 or G-protein tDNA mutants measured by RT-qPCR. **c**, Protection of 4-week-old wild-type Col-0 or *gαβ* double mutant leaves from *P. syringae* pv. *tomato* strain DC3000 infection by pre-infiltrated PrpL or flg22; *gpa1-4* is a *gα* mutant, *agb1-2* is a *gβ* mutant, and *gpa1-4/agb1-2* is a *gαβ* double mutant. Data represent mean ± s.d.; $n = 3$ biological replicates with each experiment containing eight seedlings (**b**) and $n = 10$ leaves from five plants (**c**); $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$, Student's *t*-test versus Col-0 (**b**) and versus mock (**c**).

affected in a *gβ* mutant than in a *gα* mutant, protection against *P. syringae* was only modestly affected in a *gαβ* double mutant, and the activation of MAPKs was not affected in any of the G-protein mutants (Fig. 2a–c and Extended Data Fig. 6b). These data show that G-protein signalling is required to activate downstream MAPKs in response to protease IV and ArgC, but not flg22 (Fig. 2a), and that G proteins play different roles in canonical microbe-associated molecular pattern molecule and protease-mediated signalling pathways.

In a search of potential signalling components that could link the heterotrimeric G-protein complex to downstream MAPK cascades, we considered the conserved scaffold protein RACK1 (ref. 7). The rationale was that RACK1 shares about 25% amino-acid sequence identity with Gβ and like Gβ has a seven-bladed β-propellor structure[7], interacts with Gβ in metazoans[8], and functions in innate immune signalling in rice[9]. There are three RACK1 homologues in *Arabidopsis*: RACK1A, 1B, and 1C, which share about 90% amino-acid sequence identity[10].

We used three methods to determine whether *Arabidopsis* RACK1 proteins interact with G proteins and MAPKs. In a bimolecular fluorescence complementation (BiFC) assay in *Nicotiana benthamiana* leaves, RACK1A, RACK1B, and RACK1C interacted with Gβ, MEKK1(K361M), MKK4, MKK5, MPK3, and MPK6, but not Gα or MPK4 (Extended Data Fig. 7a). The kinase-inactive version of MEKK1, MEKK1(K361M), was used in this experiment because the auto-activation of native MEKK1 destabilizes its interaction with RACK1 (data not shown). MEKK1, MKK4/5, and MPK3/6 are the *Arabidopsis* MAPK kinase kinase (MAPKKK), MAPK kinases (MAPKKs), and MAPKs, respectively, that were proposed to constitute an MAPK-signalling cascade in the flg22/FLS2 signalling pathway[11]. Similar results to those obtained with the BiFC assay in *N. benthamiana* were obtained with BiFC and split firefly luciferase complementation (SFLC) assays for RACK1A interactors in *Arabidopsis* protoplasts (Extended Data Fig. 7b, c). The interaction between RACK1 proteins and MPK3/6, but not MPK4, is consistent with the data in Fig. 1b, showing that MPK6 and MPK3, but not MPK4, are strongly activated after protease IV treatment.

In co-immunoprecipitation experiments in *Arabidopsis* mesophyll protoplasts using Flag-tagged RACK1 proteins as the bait and HA-tagged

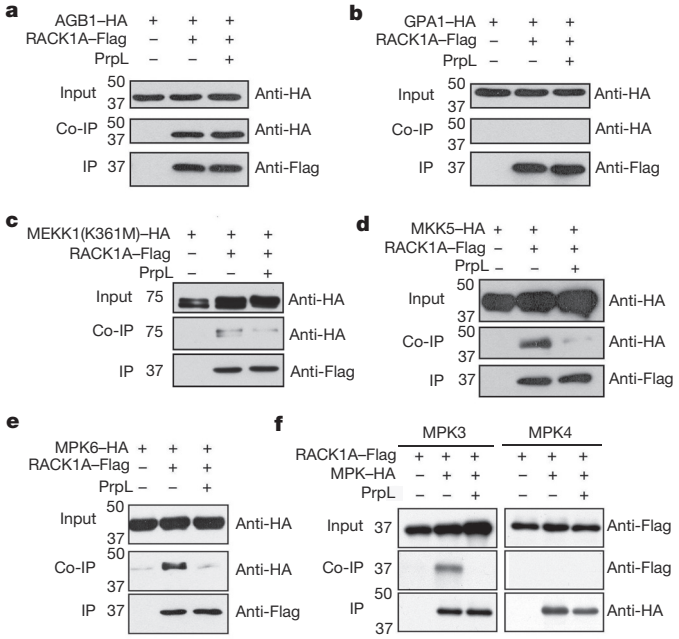


Figure 3 | RACK1A interacts with Gβ and MAPKs. a–f, Co-immunoprecipitation (Co-IP) assays in *Arabidopsis* protoplasts. Protoplasts were treated with 100 nM purified PrpL for 15 min. Target proteins were detected in western blots using anti-HA or anti-Flag antibodies. Numbers on the left axis of blots represent marker size (molecular mass in kilodaltons).

Gβ subunit as the prey, we observed binding between all three *Arabidopsis* RACK1 proteins and Gβ (Fig. 3a and Extended Data Fig. 7d). In contrast to Gβ, HA-tagged Gα was not pulled down by Flag-tagged RACK1 proteins (Fig. 3b and Extended Data Fig. 7d). In the co-immunoprecipitation experiments, the interaction of Gβ with RACK1A was not dependent on Gα, because the interaction was still present in the *gαβ* double mutant (Extended Data Fig. 7e). Finally, consistent with the BiFC and SFLC assays, HA-tagged MEKK1(K361M), MKK5, MPK3, and MPK6 all co-immunoprecipitated with Flag-tagged RACK1A, whereas MPK4 did not under the same condition (Fig. 3c–f). The amounts of the MAPKK and MAPKs that were pulled down by RACK1A in the co-immunoprecipitation experiments clearly decreased in the presence of protease IV (Fig. 3d–f), suggesting that protease IV releases the activated MAPKs from the RACK1–MAPK cascade complex to execute their downstream cellular functions. In the case of the MAPKKK MEKK1, we also identified endogenous RACK1 proteins by mass spectrometric analysis as binding partners of MEKK1(K361M) (Extended Data Fig. 7f) in a transgenic line in which Flag-tagged MEKK1(K361M) is expressed under the control of the 3.9-kilobase (kb) *MEKK1* native promoter in a *mekk1* null mutant background.

To confirm the physiological relevance of the observed interactions between RACK1 and MAPK cascade components (Fig. 3 and Extended Data Fig. 7), we tested a variety of loss-of-function MAPK mutants and knockdowns. We found that the activation of the defence-related genes *WRKY30* and *WRKY33* by protease IV was almost completely blocked in two independent *mpk3,6-es* transgenic lines in which *mpk3* is silenced with an oestradiol-inducible *MPK3*-RNA interference (RNAi) construct in a null *mpk6* mutant background (Extended Data Fig. 8a). We also found that both protease-IV-triggered MPK3/6 activation and *WRKY30* and *WRKY33* gene induction were disrupted in *mkk4,5-es* transgenic lines (Extended Data Fig. 8a, b), which utilize a single oestradiol-inducible RNAi construct to target both *MKK4* and *MKK5* messenger RNAs (mRNAs). Finally, we observed a significant decrease in protease-IV-triggered induction of *WRKY30* and *WRKY33* mRNA accumulation in two *mekk1* mutants, an *mekk1* null mutant, and the *mekk1* null mutant complemented with an MEKK1(K361M) construct (*mekk1/pMEKK1::MEKK1(K361M)*) (Extended Data Fig. 8c, d). As previously reported, MEKK1(K361M), which is deficient in kinase activity, rescues the severe growth defect of an *mekk1* null mutant[12]. In contrast to the *mkk4,5* knockdown lines, we did not consistently observe a decreased level of protease-IV-triggered MPK3/6 phosphorylation in either of the *mekk1* mutants (Extended Data Fig. 8e). One explanation for the partial decrease in *WRKY* gene induction but not in MPK3/6 phosphorylation in the *mekk1* mutants is that multiple MAPKKKs[13] function additively to activate MPK3/6 but that the phosphorylation assay is not sensitive enough to detect a partial loss of MAPKKK activity.

Obtaining genetic evidence that RACK1 is required for protease-mediated signalling is challenging because of the functional redundancy of the three RACK1 proteins in *Arabidopsis*. Transfer-DNA (tDNA) mutants corresponding to insertions in individual *rack1* genes did not show any decrease in protease-IV- or flg22-activated MAPK levels (Extended Data Fig. 9a), and only moderate decreases in protease-IV- but not flg22-triggered defence gene induction (Extended Data Fig. 9b). Because *rack1a rack1b rack1c* triple null mutants have a dwarf phenotype and do not set seeds[14], we generated two independent transgenic lines, *amiR–rack1–es1* and *amiR–rack1–es2*, which express a previously described artificial microRNA (amiR–*RACK1*-4)[15] under the control of an oestradiol-inducible promoter. These transgenic lines showed dramatically decreased transcript levels of all three *rack1* genes following oestradiol treatment (Extended Data Fig. 9c). Following protease IV or ArgC treatment, *amiR–rack1–es1* and *amiR–rack1–es2* seedlings that had been induced with oestradiol exhibited markedly decreased levels of activated MPK3 and MPK6 (Fig. 4a). Protoplasts transfected with constitutively expressed amiR–*RACK1*-4 also showed reduced levels of protease-IV-mediated MPK3 and MPK6 activation (Extended Data

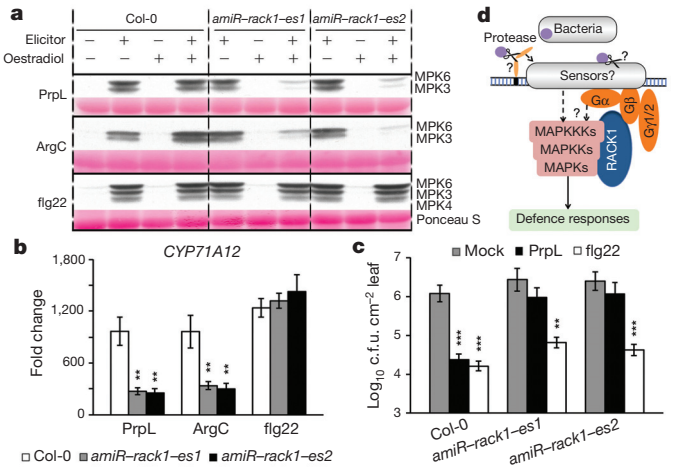


Figure 4 | Transiently silencing all three *rack1* genes abrogates proteases but not flg22-mediated responses. a, Three-day-old wild-type Col-0 and transgenic *Arabidopsis* seedlings from two independent *amiR–rack1–es* lines were treated with oestradiol to activate expression of the artificial microRNA constructs and then 2 days later were treated with PrpL, ArgC or flg22 and harvested for the MAPK phosphorylation assay. The same molecular mass region of the western blot is shown as in Fig. 1b. **b**, Seedlings were treated with oestradiol followed by PrpL, ArgC or flg22 as in panel **a** and then harvested for RT-qPCR analysis of *CYP71A12* transcript levels. Water-treated Col-0 was used as a normalization control. **c**, Protection of 4-week-old wild-type Col-0 and transgenic *amiR–rack1–es1* or *amiR–rack1–es2* plants from *P. syringae* pv. *tomato* strain DC3000 infection mediated by PrpL or flg22 24 h after treatment with oestradiol. Data represent mean ± s.d.; $n = 3$ biological replicates with each experiment containing 12 seedlings (**b**) and $n = 10$ leaves from five plants (**c**); $^{**}P < 0.01$; $^{***}P < 0.001$, Student's *t*-test versus Col-0 (**b**) and versus mock (**c**). **d**, A model of protease-activated novel innate immune signalling pathway in *Arabidopsis*.

Fig. 9d, e). Similarly, knockdown of the *rack1* genes blocked protease-IV- or ArgC-mediated defence gene induction (Fig. 4b) and protease-IV-mediated protection against *P. syringae* infection (Fig. 4c). In contrast to protease IV and ArgC, flg22-mediated activation of MAPKs or defence gene expression or protection against *P. syringae* were not affected by knockdown of the *rack1* genes (Fig. 4a–c and Extended Data Fig. 9e). These data are consistent with the conclusion that RACK1 proteins function in the protease IV and ArgC signalling pathway but not the flg22 pathway.

The RACK1 proteins studied here are the first MAPK cascade scaffolding proteins discovered for the large family of plant genes encoding MAPK cascade components. In yeast, the scaffolding protein Ste5 links an MAPK cascade to G-protein signalling in the mating pathway that is mediated by G-protein-coupled receptor stimulation by yeast pheromone[16]. In mammals, the scaffolding protein β-arrestin 2 brings MAPK cascade activity under the control of upstream G-protein-coupled receptors[16]. However, since plants do not have canonical G-protein-coupled receptors or orthologues of Ste5 and β-arrestin[6,16], our data suggest that the linkage of G-proteins to MAPKs via RACK1 is mechanistically distinct from G-protein signalling in metazoans and yeast.

The protease-activated signalling pathway is summarized in the model shown in Fig. 4d. It remains to be determined whether the cleavage of protein targets by protease IV directly or indirectly activates downstream responses. In the latter possibility, pathogen-secreted proteases could release host polypeptides that function as damage-associated molecular patterns which are subsequently recognized by corresponding immune receptors. In either case, an evolutionary and physiological interpretation of our findings is that plants evolved a new surveillance system to recognize and respond to pathogen-encoded proteases that disrupt host homeostasis via their proteolytic activity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 June 2014; accepted 15 January 2015.
Published online 2 March 2015.

1. Tena, G., Boudsocq, M. & Sheen, J. Protein kinase signaling networks in plant innate immunity. *Curr. Opin. Plant Biol.* **14,** 519–529 (2011).
2. Arthur, J. S. & Ley, S. C. Mitogen-activated protein kinases in innate immunity. *Nature Rev. Immunol.* **13,** 679–692 (2013).
3. Millet, Y. A. *et al.* Innate immune responses activated in *Arabidopsis* roots by microbe-associated molecular patterns. *Plant Cell* **22,** 973–990 (2010).
4. Denoux, C. *et al.* Activation of defense response pathways by OGs and flg22 elicitors in *Arabidopsis* seedlings. *Mol. Plant* **1,** 423–445 (2008).
5. Traidej, M., Marquart, M. E., Caballero, A. R., Thibodeaux, B. A. & O'Callaghan, R. J. Identification of the active site residues of *Pseudomonas aeruginosa* protease IV. *J. Biol. Chem.* **278,** 2549–2553 (2003).
6. Urano, D., Chen, J.-G., Botella, J. R. & Jones, A. M. Heterotrimeric G protein signalling in the plant kingdom. *Open Biol.* **3,** 120186 (2013).
7. Ullah, H. *et al.* Structure of a signal transduction regulator, RACK1, from *Arabidopsis thaliana*. *Protein Sci.* **17,** 1771–1780 (2008).
8. Dell, E. J. *et al.* The β_γ subunit of heterotrimeric G proteins interacts with RACK1 and two other WD repeat proteins. *J. Biol. Chem.* **277,** 49888–49895 (2002).
9. Nakashima, A. *et al.* RACK1 functions in rice innate immunity by interacting with the Rac1 immune complex. *Plant Cell* **20,** 2265–2279 (2008).
10. Chen, J.-G. *et al.* RACK1 mediates multiple hormone responsiveness and developmental processes in *Arabidopsis*. *J. Exp. Bot.* **57,** 2697–2708 (2006).
11. Asai, T. *et al.* MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* **415,** 977–983 (2002).
12. Suarez-Rodriguez, M. C. *et al.* MEKK1 is required for flg22-induced MPK4 activation in *Arabidopsis* plants. *Plant Physiol.* **143,** 661–669 (2007).
13. MAPK Group (Ichimura, K. *et al.*). Mitogen-activated protein kinase cascades in plants: a new nomenclature. *Trends Plant Sci.* **7,** 301–308 (2002).
14. Guo, J. & Chen, J.-G. *RACK1* genes regulate plant development with unequal genetic redundancy in *Arabidopsis*. *BMC Plant Biol.* **8,** 108 (2008).
15. Li, J.-F. *et al.* Comprehensive protein-based artificial microRNA screens for effective gene silencing in plants. *Plant Cell* **25,** 1507–1522 (2013).
16. Witzel, F., Maddison, L. & Blüthgen, N. How scaffolds shape MAPK signaling: what we know and opportunities for systems approaches. *Front. Physiol.* **3,** 475 (2012).

Acknowledgements We thank G. Tena for generating the *mekk1/pMEKK1::MEKK1(K361M)* transgenic line, Y. Zhang for the *summ1-1* mutant, M. C. Suarez-Rodriguez and P. J. Krysan for discussion, the *Arabidopsis* Biological Resource Center for tDNA insertion lines, and M. Curtis and U. Grossniklaus for the oestradiol-inducible binary vector. We thank S. Lory for *P. aeruginosa* PAO ADD1976, and M. B. Mudgett for pVSP61. We thank N. Clay, X. Dong, S. Somerville, and Ausubel laboratory members for reading the manuscript. This work was supported by Natural Sciences and Engineering Research Council of Canada and Banting Postdoctoral Fellowships awarded to Z.C., National Science Foundation grants MCB-0519898 and IOS-0929226 and National Institutes of Health grants R37-GM48707 and P30 DK040561 to F.M.A., and National Science Foundation grant IOS-0618292 and National Institutes of Health grant R01-GM70567 to J.S.

Author Contributions Z.C., J.-F.L., J.S., and F.M.A. designed experiments, Z.C., J.-F.L., Y.N., X.-C.Z., O.Z.W., Y.X., S.D., Y.M., and J.B. performed experiments, Z.C., J.-F.L., B.J.M., J.S., and F.M.A. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.M.A. (ausubel@molbio.mgh.harvard.edu).

METHODS

No statistical methods were used to predetermine sample size.

Bacterial strains. *P. aeruginosa* strains used in this work were wild type and mutants of UCBPP-PA14 (refs 17, 18) and PAO ADD1976 (ref. 19). The latter strain carries the chromosomally incorporated gene for T7 RNA polymerase under the control of the *lac* repressor and was used for production of His-tagged PrpL and His-tagged ArgC. *Xanthomonas campestris* strains were described previously[20].

A PA14/Δ*prpL* in-frame deletion mutant was constructed using a method described previously[21] that employed a sequence that contained regions immediately flanking the coding sequence of the *prpL* gene. This fragment was generated by a standard three-step PCR protocol using Phusion DNA polymerase (New England Biolabs) and then cloned into the BamHI and HindIII sites of pEX18Ap[22], resulting in plasmid pEX18-Δ*prpL*. Plasmid pEX18-Δ*prpL* was used to introduce the deleted region into the wild-type PA14 genome by homologous recombination. *Escherichia coli* strain SM10 λpir was used for triparental mating[23].

For the purification of His-tagged protease IV or ArgC, the *P. aeruginosa* PA14 *prpL* gene or the *X. campestris* strain 1946 *argC* gene were cloned into the EcoRI and XhoI sites of pETP30 (ref. 24), creating plasmids pETP-*prpL* or pETP-*argC*, which encode 6× His-tagged PrpL and 6× His-tagged ArgC, respectively. The resulting plasmids were transformed into *P. aeruginosa* PAO ADD1976 by electroporation[25] to generate the strains ADD/pETP-*prpL* or ADD/pETP-*argC* for purification of His-tagged protease IV or His-tagged ArgC, respectively.

For *argC* complementation in *Xanthomonas*, the *X. campestris* strain 1946 *argC* gene was cloned into the BamHI site of pVSP61 (ref. 26), creating plasmid pVSP61-*argC*. An HA-tag was incorporated at the carboxy (C)-terminal of the *argC* gene to detect the complemented protein. The resulting plasmid and empty pVSP61 vector were transformed into *X. campestris* strain 8004 by triparental conjugation[23].

Antibiotics were supplemented as needed: ampicillin or carbenicillin, 50 μg ml^{-1} for *E. coli* or 300 μg ml^{-1} for *P. aeruginosa*; kanamycin 50 μg ml^{-1} for *E. coli* and *Xanthomonas campestris* or 200 μg ml^{-1} for *P. aeruginosa*; and rifampicin 100 μg ml^{-1}.

Construction of *Arabidopsis* transgenic lines. Construction of *amiR–rack1–es* transgenic lines and the *mekk1/pMEKK1::MEKK1(K361M)* transgenic line was performed as follows: the BamHI/PstI fragment of pre-amiR–*RACK1*–4 (ref. 15) was inserted between the oestradiol-inducible promoter[27] and the *NOS* terminator in a modified pUC119-RCS vector[28]. The pre-amiR–*RACK1*–4 expression cassette was then cut out by AscI digestion and inserted into AscI-digested binary vector pFGC19-XVE-RCS[28], which expresses the XVE transcriptional activator[29] under the 35S promoter, to obtain pFGC–EST–RACK1. This latter plasmid was introduced into *Agrobacterium tumefaciens* GV3101 cells by electroporation, and GV3101/pFGC–EST–RACK1 was used to generate transgenic *Arabidopsis* plants with inducible *amiR–RACK1* expression using the floral dip technique[30]. To generate *mekk1/pMEKK1::MEKK1(K361M)* transgenic *Arabidopsis*, an ~9.4 kb *MEKK1* genomic fragment was used to complement a *mekk1* null mutant (Salk_052557). This genomic fragment contains an ~3.9 kb promoter sequence upstream of the start codon, an 'AAGG' to 'ATGG' mutation in exon 2 (corresponding to K361M mutation in MEKK1) to disrupt MEKK1 kinase activity, and a double Flag-tag coding sequence upstream of the stop codon.

Fractionation of the PA14 secretome. One litre of PA14 cells grown in M9 minimal medium (6.8 g l^{-1} Na_2HPO_4, 3 g l^{-1} KH_2PO_4, 0.5 g l^{-1} NaCl, 1 g l^{-1} NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 10 μM $FeCl_3$, 0.4% glucose, 10 mg l^{-1} thiamine) was centrifuged at 20,000*g* at 4 °C for 30 min and the pellet was discarded. The supernatant was filtered through a 0.22 μm low protein-binding filter (Corning). Secreted PA14 proteins in the filtrate were precipitated with ammonium sulphate (85% saturation) at 4 °C overnight, followed by centrifugation at 20,000*g* at 4 °C for 1 h. The pellet was resuspended in 30 ml buffer A (20 mM Tris, pH 8.8), concentrated to 150 μl using Centrion Plus-70 filter (Millipore) to remove the excess ammonium sulphate, and diluted again into 10 ml buffer A. The protein sample was loaded onto a 1-mL DEAE anion-exchange chromatography column (GE Healthcare) that was washed with buffer B (20 mM Tris, pH 8.8, 1 M NaCl) and equilibrated with buffer A. Proteins were separated into 1-ml fractions with a linear gradient of buffer B (0–60% within 20-column volumes). The fractionation was performed at 4 °C with a flow rate of 1 ml min^{-1}.

Purification of *P. aeruginosa* protease IV and *X. campestris* ArgC. Secreted proteins from ADD1976/pETP-*prpL* were precipitated as described above and resuspended in lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0). The sample was loaded onto a 5-mL HisTrap Affinity Column (GE Healthcare) and the 6× His tagged PA14 protease IV was purified according to the manufacturer's instructions. The eluted protease IV was concentrated to 150 μl and immediately subjected to a Superdex 200 gel filtration column (GE Healthcare). Purified protease IV was exchanged into M9 minimal medium and filter-sterilized using a 0.22 μm low protein-binding filter (Millipore). The concentration of the purified protease IV was adjusted to 20 μM, aliquoted, and stored at −80 °C before being used for plant treatments. *X. campestris* protease ArgC was purified using the same protocol.

Protease assay. The protease activity assay of protease IV and its homologue ArgC was determined as previously described[31]. Protease IV and ArgC were inactivated by TLCK as previously described[31].

Plant growth. Seeds were sterilized in 20% bleach for 2 min and washed three times with sterile water. Seedlings were grown in liquid MS medium (Murashige and Skoog basal medium with vitamins from Phytotechnology Laboratories supplemented with 0.5 g l^{-1} MES hydrate and 0.5% sucrose at pH 5.7) in either 24-well assay plates (BD Falcon) (eight seeds and 0.5 ml medium per well) for MAPK assays, microarray and RT-qPCR analysis, callose induction and GUS expression, or 96-well plates (Greiner Bio-One) (one seed and 0.2 ml medium per well) for oxidative burst measurements. Plates were sealed with Micropore tape and placed on grid-like shelves over water trays on a Floralight cart in a plant growth chamber for 10 days at 21 °C with 75% relative humidity under 16 h of daylight (65–70 μE m^{-2} s^{-1}). The media in 24-well plates was exchanged for fresh media on day 8, whereas the media in 96-well plates was exchanged for sterile water on day 9.

Elicitor treatments. The synthetic peptide flg22 was synthesized by Genscript. Experimentally determined optimal concentrations of protease IV were as follows: 20 nM for oxidative burst measurements, microarray and RT-qPCR analyses; 40 nM for MAPK activation; 100 nM for GUS expression; 500 nM for callose elicitation and the infection protection assay. For direct comparison, the same concentrations of flg22 and protease IV or ArgC were used in the same assays. Ten-day-old seedlings were treated with different elicitors for the following times unless otherwise specified: 6 h for GUS assays in reporter line *CYP71A12pro:GUS*; 10 min for MAPK activation assays; 1 h or 6 h for RT-qPCR analysis of selected genes; and 18 h for callose induction.

Transient silencing of *MAPK or MAPKK* genes in transgenic plants. In two independent *mpk3,6-es* transgenic lines, *MPK3* was silenced with an oestradiol-inducible *MPK3*-RNAi construct in a null *mpk6* mutant (Salk_062471) background. In two *mkk4,5-es* transgenic lines, a single oestradiol-inducible RNAi construct was used to target both *MKK4* and *MKK5* mRNAs. Details of the construction of the *mpk3,6-es* and *mkk4,5-es* transgenic lines will be described elsewhere. The transgenic and control plants were grown in MS medium in a 24-well plate as described above for 4 days. Then the medium was changed to MS medium containing 10 μM oestradiol (Sigma, 100 mM stock in dimethylsulphoxide (DMSO)). After exposure to oestradiol for 3 days, the seedlings were treated with water and 40 nM purified protease IV for 10 min (for MAPK assays) or 20 nM purified protease IV for 1 h (for RT-qPCR assays).

Transient silencing of *rack1* genes in protoplasts and transgenic plants. Mesophyll protoplasts isolated from leaves of 4-week-old *Arabidopsis* plants (4×10^4 cells in 200 μl) were transfected with 40 μg (20 μl) of amiR–*RACK1*-4 construct or empty artificial microRNA expression vector[15] as a control. After 24 h of expression, 100 nM flg22 or 100 nM purified protease IV was added to the protoplasts followed by incubation for 10 min before the cells were harvested for MAPK assays and *rack1* gene silencing confirmation by RT-qPCR.

For oestradiol-induced *rack1* silencing in transgenic *amiR–rack1–es* lines, the wild-type Col-0 and transgenic plants were grown in MS medium in a 24-well plate as described above for 3 days. Then the medium was changed to MS medium containing 10 μM oestradiol (Sigma, 100 mM stock in DMSO). After exposure to oestradiol for 2 days, the seedlings were treated with water and 40 nM flg22 or 40 nM purified protease IV for 10 min (MAPK assay) or 20 nM flg22 or 20 nM purified protease IV for 6 h (RT-qPCR measuring transcript levels of *CYP71A12*, *GST6*, and the three *rack1* genes). For the protease-IV-mediated protection assay against *P. syringae* DC3000, 20 μM oestradiol was infiltrated into 4-week-old control Col-0 and transgenic *amiR–rack1–es1* and *amiR–rack1–es2* leaves 24 h before the mock treatment or treatment with 500 nM purified protease IV.

Mutant seed stocks. Transfer-DNA insertion lines *gpa1-4* (CS6534), *agb1-2* (CS6536), *agg1-1c* (CS16550), *agg2-1* (SALK_022447), *gpa1-4/agb1-2* (CS6535), *agg1-1c/agg2-1* (CS16551), *mekk1* (SALK_052557), *rack1a-3* (CS862351), *rack1b-2* (SALK_145920), *rack1b-3* (CS863092), *rack1c-2* (SALK_017913), and *rack1c-3* (SALK_001973) were obtained from the *Arabidopsis* Biological Resource Center.

GUS histochemical assay. After treatment with 100 nM flg22 or 100 nM purified protease IV for 6 h, plants were washed with 50 mM sodium phosphate (pH 7) and 0.5 ml of GUS substrate solution (50 mM sodium phosphate, pH 7, 10 mM EDTA, 0.5 mM $K_4[Fe(CN)_6]$, 0.5 mM $K_3[Fe(CN)_6]$, 0.5 mM X-Gluc, and 0.1% v/v Triton X-100) was added to each well. The plants were vacuum-infiltrated for 5 min and then incubated at 37 °C for 4 h. Tissues were fixed with a 3:1 ethanol:acetic acid solution at 4 °C overnight and placed in 95% ethanol. Tissues were cleared in lactic acid and then examined using a Discovery V12 microscope (Zeiss). For the screen of PA14 secretome fractions, 100 μl of buffer A (20 mM Tris, pH 8.8) or different DEAE fractions were added to each well.

MAPK activity. Total proteins in seedling or protoplast lysates were resolved on a 10% SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gel and transferred to a polyvinylidene difluoride membrane. Western blot analysis was conducted by using anti-phospho ERK antibodies (Cell Signaling) as the primary antibody at 1:10,000 dilution in 5% BSA and horseradish peroxidase (HRP)-conjugated anti-rabbit antibodies as the secondary antibody at 1:10,000 dilution in 5% non-fat milk. The immunoblot signal was visualized with a SuperSignal West Femto kit (Thermo Scientific).

Oxidative burst measurement. H_2O_2 was detected using a luminol–HRP-based chemiluminescence assay. A 10 mg ml^{-1} 500× HRP (Sigma-Aldrich) stock solution was prepared by dissolving 10 mg HRP in water. A 20 mg ml^{-1} 500× luminol (Sigma-Aldrich) stock solution was prepared by dissolving 20 mg luminol in 100 mM KOH. For each elicitor, a master reaction mixture was prepared by diluting individual elicitor, HRP, and luminol stocks with water. The plates were kept in the dark for 1 h before elicitation. The following procedures were performed in the dark. Liquid was removed at the end of the 1-h pre-treatment and 200 µl of master reaction mixture was added into each well. Plates were placed into a 96-well scintillation reader immediately and light emission was monitored using a 96-well scintillation counter (1450 Microbeta Wallac TriLux Scintillation/Luminescence counter). Every plate was read for about 30 cycles. Kinetics of H_2O_2 production was determined by plotting the average chemiluminescence counts from all the seedlings under the same condition over the reading period. Every time point is the mean value of 16 seedlings.

RNA isolation and microarray and RT-qPCR analysis. Total RNA was isolated according to the manufacturer's instructions using an RNeasy Plant Mini Kit (Qiagen). DNA was removed using the DNA-free kit (Ambion), and reverse transcription reactions were performed using an iScript cDNA synthesis kit (Bio-Rad). Complementary DNA (cDNA) concentrations were measured using a Nano-drop instrument (Thermo Scientific). RT-qPCR was performed using a CFX96 real-time PCR machine (Bio-Rad) using iQ SYBR Green Supermix (Bio-Rad). The following PCR reaction programme was used: 95 °C for 3 min followed by 50 cycles of 95 °C for 30 s and 55 °C for 30 s. Fold change was calculated relative to plants treated with M9 buffer. Fold induction data represent the mean ± s.d., $n = 3$ with each containing eight seedlings. Expression values were normalized to that of the eukaryotic translation initiation factor 4A1 (*EIF4A1*). The primers used were the following: *EIF4A1* (At3g13920), 5′-GCAGTCTCTTCGTGCTGACA-3′ and 5′-TGTCATAGATCTGGTCCTTGAA-3′; *CYP71A12* (At2g30750), 5′-GATTATCACCTCGGTTCCT-3′ and 5′-CCACTAATACTTCCCAGATTA-3′; *WRKY30* (At5g24110), 5′-GCAGCTTGAGAGCAAGAATG-3′ and 5′-AGCCAAATTTCCAAGAGGAT-3′; *GST6* (At2g47730), 5′-CCATCTTCAAAGGCTGGAAC-3′ and 5′-TCGAGCTCAAAGATGGTGAA-3′; *WRKY29* (At4g23550), 5′-ATCCAACGGATCAAGAGCTG-3′ and 5′-GCGTCCGACAACAGATTCTC-3′; *WRKY33* (At2g38470), 5′-GGGAAACCCAAATCCAAGA-3′ and 5′-GTTTCCCTTCGTAGGTTGTGA-3′; *ERF1* (At3g23240), 5′-TCGGCGATTCTCAATTTTTC-3′ and 5′-ACAACCGGAGAACAACCATC-3′; *rack1a* (At1g18080), 5′-GCTGAAAAGGCTGACAACAGT-3′ and 5′-GCTCCAGTTAAGGCTTGTGC-3′; *rack1b* (At1g48630), 5′-TTGTTGAGGATTTGAAGGTTGA-3′ and 5′-CCAGTTCAAGCTTGTGCAGTA-3′; *rack1c* (At3g18130), 5′-GAGGCAGAGAAGAATGAAGGTG-3′ and 5′-CCAGTTCAAGCTTGTGCAGTA-3′. *WRKY* gene induction was measured 1 h after elicitation, whereas *CYP71A12*, *ERF1*, and *GST6*, were measured 6 h after elicitor treatment.

For microarray analysis, RNA quality was assessed by checking the integrity of RNA on an Agilent 2100 Bioanalyzer (Agilent Technologies). Target labelling was performed according to the protocol given in the Affymetrix GeneChip 3′ IVT Express Kit Technical Manual. Microarray hybridizations and scanning were finished at the Genomics Core, Joslin Diabetes Center, Boston, Massachusetts. Microarray CEL files were read into the R statistical analysis software, version 2.15.2. Arrays were analysed together using the standard robust multi-array average procedure as implemented in Bioconductor's 'affy' package, version 1.36.1 (refs 32, 33). Fold changes were calculated using $\log_2$-transformed expression values by subtracting the mean of control samples from the mean of treated samples. Microarray CEL files were also obtained from previous studies exploring the effects of flg22 and oligogalacturonides on gene expression[4]. These two experiments were subjected to the robust multi-array average procedure together, but downstream analyses (for example, fold change computations) were performed separately on the two treatments. The microarray data have been deposited in the GEO database under accession number GSE58518.

Callose deposition assay. Elicitor-induced callose deposition in cotyledons of 10-day-old *Arabidopsis* seedlings was detected using aniline blue as described[34]. Eighteen hours after elicitation, seedlings were fixed under a vacuum in 3:1 ethanol:acetic acid. The clearing solution was changed until the leaves were colourless. Tissues were washed in 70% ethanol and then 50% ethanol for at least 2 h each time and rehydrated in several brief H_2O washes followed by an overnight H_2O wash. Samples were then made transparent by several minutes in a vacuum with 10% NaOH followed by a 2-h incubation at 37 °C on a shaking platform. After several more H_2O washes, tissues were incubated in the dark at 21 °C for at least 4 h with 0.01% aniline blue in 150 mM K_2HPO_4 (pH 9.5). After mounting on slides in 50% glycerol, samples were examined with a Zeiss Axioplan microscope using ultraviolet illumination and a broadband 4′,6-diamidino-2-phenylindole (DAPI) filter set (excitation filter 390 nm; dichroic mirror 420 nm; emission filter 460 nm).

Pathogenicity assays. *Arabidopsis* pathogenicity assays, including infection by *P. syringae* strain DC3000 with or without pre-infiltration of protease IV or flg22, were performed according to previously described protocols[21]. Data represent the mean of bacterial titres ± s.d. of ten leaf disks excised from ten leaves of five plants. The infection protection assay was repeated three times with similar results.

Xanthomonas pathogenicity assays in *B. oleracea* were performed according to previously described protocols[35] with modifications. Seeds of broccoli cultivar *B. oleracea* var. Marathon were sown in Fafard number 2 soil mix and grown in a 12-h light (70 µE m^{-2} s^{-1}) cycle at 19 °C and 60% relative humidity. Individual seedlings were transferred to 5 cm × 5 cm pots after one week and kept at a cycle of 16-h light (150 µE m^{-2} s^{-1}) at 23 °C followed by 8-h dark at 20 °C and 70% relative humidity. After a further 2 weeks of growth, the 3-week-old plants were used for *Xanthomonas* infiltration. Fresh *X. campestris* overnight cultures were washed and adjusted to 10^6 cells per millilitre in 10 mM $MgSO_4$. A standard infiltration protocol was used to infect 3-week-old leaves. After infection, the plants were transferred to a growth chamber with the following conditions: 12-h light (60 µE m^{-2} s^{-1}) at 28 °C at 90% relative humidity for 2 days before being harvested for counting of colony-forming units. Data represent the mean of bacterial titres ± s.d. of ten leaf disks excised from ten leaves of five plants. The infection protection assay was repeated three times with similar results.

Co-immunoprecipitation. For co-immunoprecipitation performed in protoplasts, mesophyll protoplast isolation from leaves of 4-week-old *Arabidopsis* plants and polyethylene glycol (PEG)-mediated DNA transfection were performed as previously described[36]. Co-immunoprecipitation was performed as described previously[37] with modifications. Briefly, 100 µg (50 µl) of PREY plasmids were used to co-transfect 1 ml *Arabidopsis* mesophyll protoplasts (5×10^5 cells) with 100 µg (50 µl) of BAIT plasmids or empty vectors. After 6 h to allow protein expression, the cells were pelleted and lysed in 200 µl of immunoprecipitation buffer (10 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA, 10% glycerol, 1% Triton X-100, 1× Roche EDTA-free protease inhibitor cocktail) by vigorous vortexing for 1 min. Twenty microlitres of lysate was saved as the input fraction to ensure that the Prey proteins were expressed equally in all samples. The rest of the lysate (180 µl) was mixed with 320 µl immunoprecipitation buffer and vigorously vortexed for 1 min. The resultant clear lysate was centrifuged at 21,000*g* for 10 min at 4 °C, and the supernatant was incubated with a 10 µl slurry of anti-Flag M2 agarose beads (Sigma) or anti-HA magnetic beads (Pierce) for 3 h at 4 °C. The beads were washed three times with the immunoprecipitation buffer and once with 50 mM Tris-HCl, pH 7.5. The eluate was obtained by boiling the beads in 40 µl of SDS–PAGE loading buffer and the presence of co-immunoprecipitated PREY proteins was detected by immunoblotting analysis using HRP-conjugated anti-HA antibody or anti-Flag (Roche) at 1:10,000 dilution; the immunoblot signal was visualized using a SuperSignal West Femto kit (Thermo Scientific). The same membrane was stripped and re-used to detect the comparable amounts of immunoprecipitated BAIT proteins by immunoblot.

BiFC. For plasmids used in the split-mCherry assay, the coding sequence of the amino (N)-terminal fragment (mCherryN, amino acids 1–159) or the C-terminal fragment (mCherryC, amino acids 160–235) of mCherry was PCR amplified, digested by BamHI/NotI, and inserted into the same digested pAN vector, which contained a double 35S promoter and a NOS terminator, to obtain pcCherryN and pcCherryC plasmids. Genes for protein–protein interaction tests were inserted into the XbaI/BamHI-digested pcCherryN or pcCherryC vectors after digestion of their PCR products with XbaI (or SpeI, NheI if the XbaI site was present in the gene) at the 5′ end and with BamHI (or BglII if the BamHI site was present in the gene) at the 3′ end, allowing the expression of a chimaeric gene of interest with the coding sequence of mCherryN or mCherryC at the 3′ end.

For binary plasmids used in the BiFC assay in agroinfiltrated *N. benthamiana* leaves, pFGC-RCS (kanamycin resistant) and pPZP-RCS (spectinomycin resistant) binary vectors were constructed by replacing the original sequences between EcoRI and HindIII of pFGC19 and pPZP222 with the multiple cloning site sequence from pUC119-RCS flanked by EcoRI and HindIII[38]. Subsequently, the entire expression cassette of 'gene'-mCherryN was PCR amplified from protoplast expression plasmids, digested by AscI and inserted into the AscI site of pFGC-RCS, while the entire expression cassette containing the 'gene'-mCherryC fusion DNA was PCR amplified from protoplast expression plasmids, digested by AscI and inserted into the AscI site of pPZP-RCS. A pair of pFGC-RCS and pPZP-RCS plasmids expressing a pair of genes for protein–protein interaction tests were co-transformed into *Agrobacterium* GV3101 cells by electroporation, and cells transformed with both binary plasmids were selected by the addition of both kanamycin and spectinomycin to

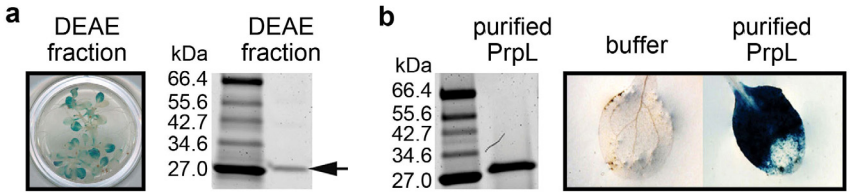
the growth medium. Leaves of 4- to 5-week-old *N. benthamiana* plants were infiltrated with agrobacteria (final attenuance, $D_{600\,nm} = 0.01$) containing constructs expressing the mCherryN fragment fused to GPA1, AGB1, or MAPKs and the mCherryC fragment fused to RACK1A/B/C. The agroinfiltration experiment was performed as described previously[39].

Arabidopsis protoplasts 18 h after transfection and *N. benthamiana* leaf pieces 2 days after agroinfiltration were imaged using a Leica DM-6000B upright fluorescence microscope with phase and differential interference contrast equipped with a Leica FW4000 digital image-acquisition and processing system.

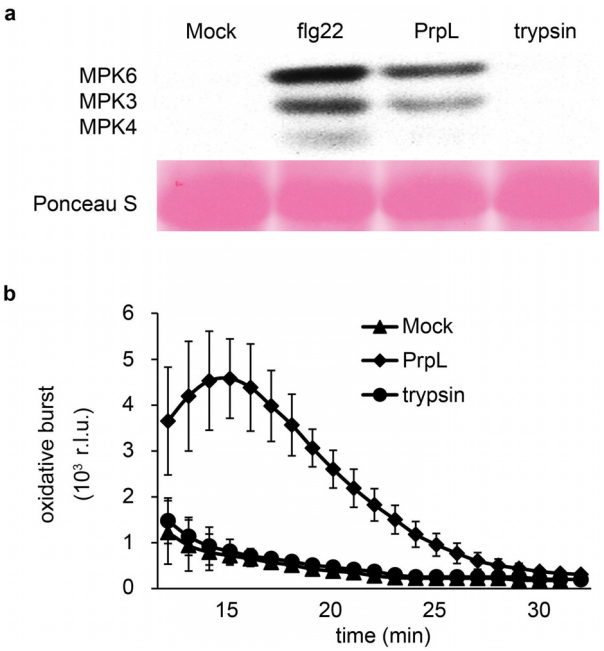
SFLC. For plasmids used in the SFLC assay, the genes for protein–protein interaction tests were inserted into the XbaI/BamHI-digested pcFLucN or pcFLucC vectors[37] after digestion of their PCR products with XbaI (or SpeI, NheI if the XbaI site was present in the gene) at the 5′ end and with BamHI (or BglII if the BamHI site was present in the gene) at the 3′ end, allowing the expression of a chimaeric gene of interest with the coding sequence of FLucN or FLucC at the 3′ end.

SFLC experiments performed in protoplasts were performed as described previously[37]. Briefly, 10 μg (5 μl) of PREY plasmids were used to co-transfect 100 μl of *Arabidopsis* mesophyll protoplasts (5×10^5 cells) with 10 μg (5 μl) of BAIT plasmids. One microgram of UBQ10::GUS plasmid was used in each transfection as an internal normalization control. After 6 h to allow for protein expression, the luminescence of each sample was recorded by a GloMax-Multi microplate multimode reader (Promega) with the integration time set as 1 s.

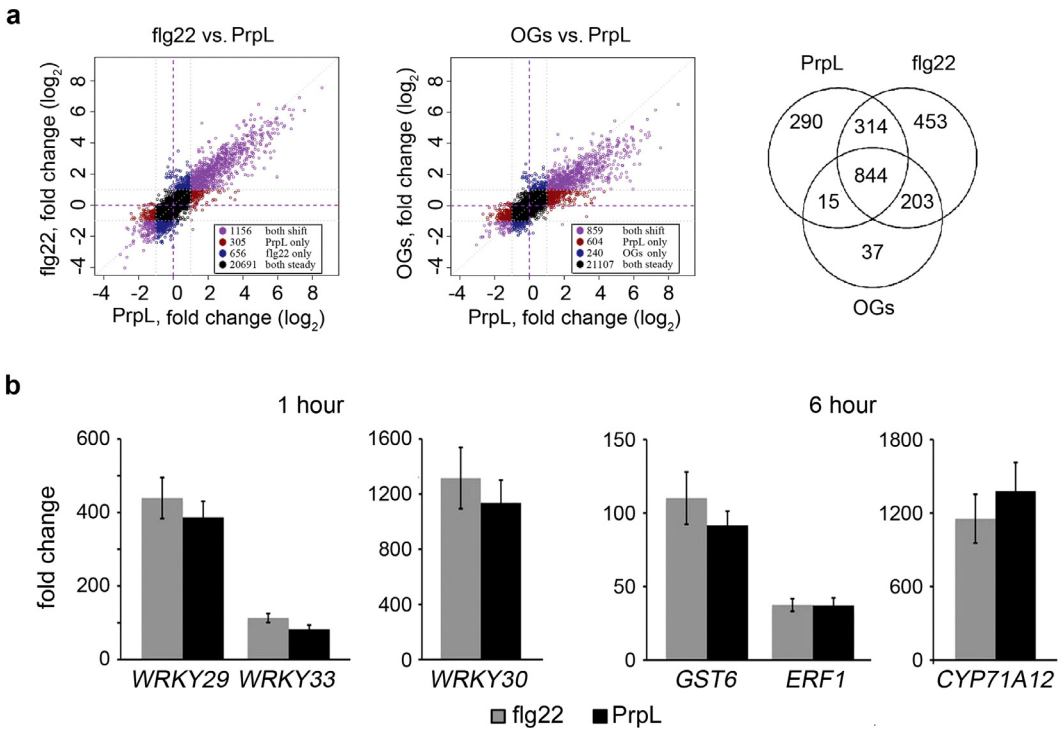
17. Rahme, L. G. *et al.* Common virulence factors for bacterial pathogenicity in plants and animals. *Science* **268,** 1899–1902 (1995).
18. Liberati, N. T. *et al.* An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl Acad. Sci. USA* **103,** 2833–2838 (2006).
19. Motley, S. T. & Lory, S. Functional characterization of a serine/threonine protein kinase of *Pseudomonas aeruginosa. Infect. Immun.* **67,** 5386–5394 (1999).
20. Parker, J. E., Barber, C. E., Mi-jiao, F. & Daniels, M. J. Interaction of *Xanthomonas campestris* with *Arabidopsis thaliana*: characterization of a gene from *X. c.* pv. *raphani* that confers avirulence to most *A. thaliana* accessions. *Mol. Plant Microbe Interact.* **6,** 216–224 (1993).
21. Djonović, S. *et al.* Trehalose biosynthesis promotes *Pseudomonas aeruginosa* pathogenicity in plants. *PLoS Pathog.* **9,** e1003217 (2013).
22. Prentki, P. & Krisch, H. M. *In vitro* insertional mutagenesis with a selectable DNA fragment. *Gene* **29,** 303–313 (1984).
23. Hirsch, A. M. *et al. Rhizobium meliloti* nodulation genes allow *Agrobacterium tumefaciens* and *Escherichia coli* to form pseudonodules on alfalfa. *J. Bacteriol.* **158,** 1133–1143 (1984).
24. Cheng, Z., Duan, J., Hao, Y., McConkey, B. J. & Glick, B. R. Identification of bacterial proteins mediating the interactions between *Pseudomonas putida* UW4 and *Brassica napus* (canola). *Mol. Plant Microbe Interact.* **22,** 686–694 (2009).
25. Smith, A. W. & Iglewski, B. H. Transformation of *Pseudomonas aeruginosa* by electroporation. *Nucleic Acids Res.* **17,** 10509 (1989).
26. Kim, J.-G. *et al. Xanthomonas* T3S effector XopN suppresses PAMP-triggered immunity and interacts with a tomato atypical receptor-like kinase and TFT1. *Plant Cell* **21,** 1305–1323 (2009).
27. Curtis, M. D. & Grossniklaus, U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol.* **133,** 462–469 (2003).
28. Lee, L.-Y., Fang, M.-J., Kuang, L.-Y. & Gelvin, S. B. Vectors for multi-color bimolecular fluorescence complementation to investigate protein-protein interactions in living plant cells. *Plant Methods* **4,** 24 (2008).
29. Zuo, J., Niu, Q. W. & Chua, N. H. Technical advance: an estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants. *Plant J.* **24,** 265–273 (2000).
30. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana. Plant J.* **16,** 735–743 (1998).
31. Engel, L. S., Hill, J. M., Caballero, A. R., Green, L. C. & O'Callaghan, R. J. Protease IV, a unique extracellular protease and virulence factor from *Pseudomonas aeruginosa. J. Biol. Chem.* **273,** 16792–16797 (1998).
32. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).
33. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
34. Clay, N. K., Adio, A. M., Denoux, C., Jander, G. & Ausubel, F. M. Glucosinolate metabolites required for an *Arabidopsis* innate immune response. *Science* **323,** 95–101 (2009).
35. Meyer, D., Lauber, E., Roby, D., Arlat, M. & Kroj, T. Optimization of pathogenicity assays to study the *Arabidopsis thaliana–Xanthomonas campestris* pv. campestris pathosystem. *Mol. Plant Pathol.* **6,** 327–333 (2005).
36. Yoo, S. D., Cho, Y. H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols* **2,** 1565–1572 (2007).
37. Li, J.-F., Bush, J., Xiong, Y., Li, L. & McCormack, M. Large-scale protein-protein interaction analysis in *Arabidopsis* mesophyll protoplasts by split firefly luciferase complementation. *PLoS ONE* **6,** e27364 (2011).
38. Li, J.-F., Park, E., von Arnim, A. G. & Nebenfuhr, A. The FAST technique: a simplified *Agrobacterium*-based transformation method for transient gene expression analysis in seedlings of *Arabidopsis* and other plant species. *Plant Methods* **5,** 6 (2009).
39. King, S. R. F. *et al. Phytophthora infestans* RXLR effector PexRD2 interacts with host MAPKKKε to suppress plant immune signaling. *Plant Cell* **26,** 1345–1359 (2014).



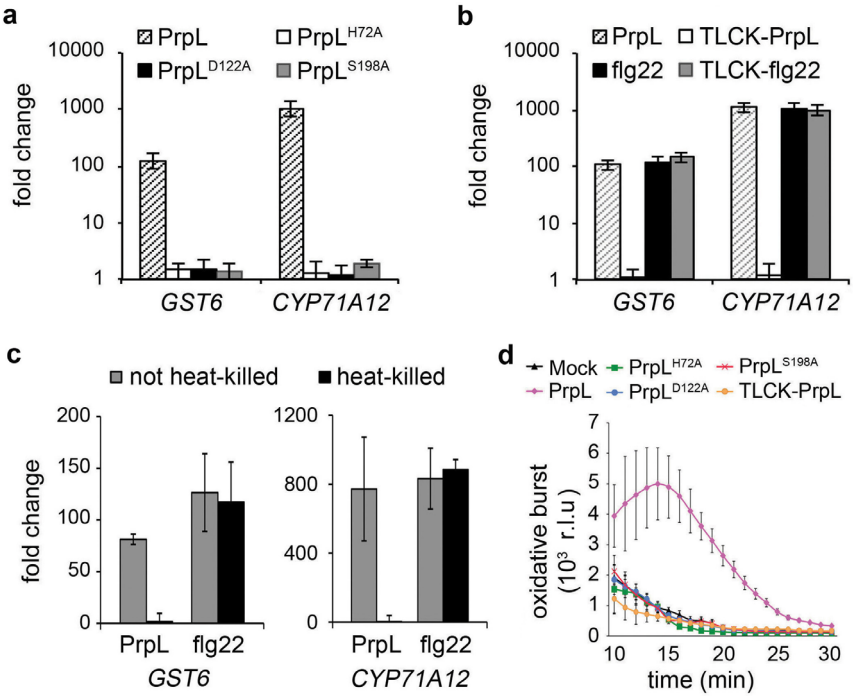
Extended Data Figure 1 | Protease IV-triggered GUS staining in *CYP71A12pro:GUS* transgenic *Arabidopsis* seedlings. **a**, Activation of *CYP71A12pro:GUS* by a DEAE fraction of the PA14 secretome (left) and purification of the eliciting activity by DEAE chromatography (right). **b**, Activation of *CYP71A12pro:GUS* in 10-day-old seedlings by 100 nM purified PrpL. The experiments in **a** and **b** were repeated three times with similar results.



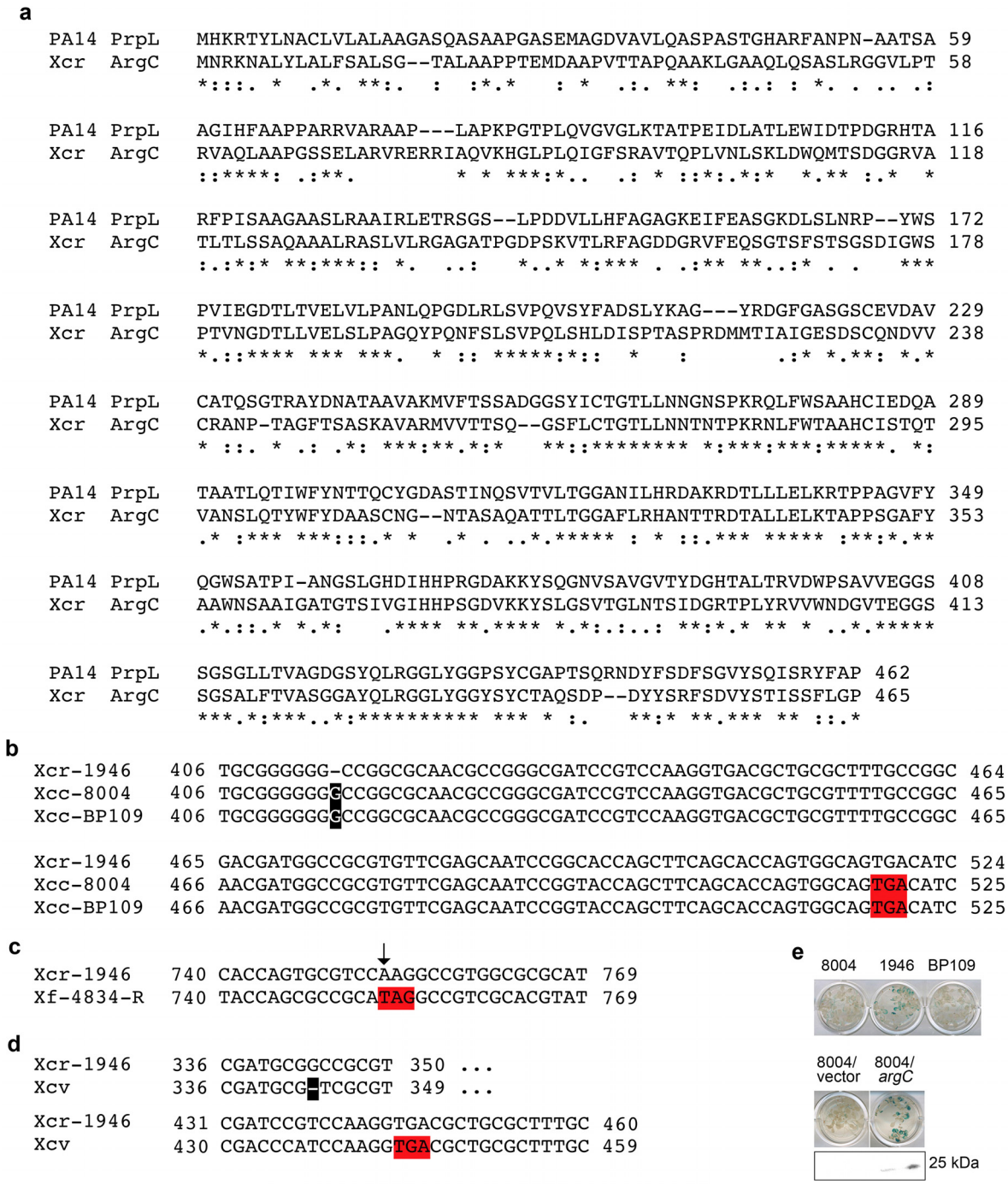
Extended Data Figure 2 | Trypsin does not activate MAPK cascade or elicit an oxidative burst in *Arabidopsis*. **a**, Western blot depicting activation of MAPKs by 40 nM flg22, or 40 nM purified PrpL, or trypsin in 10-day-old seedlings. The same molecular mass region of the western blot is shown as in Fig. 1b. **b**, Chemiluminescence assay showing elicitation of an oxidative burst in 10-day-old seedlings by 20 nM purified PrpL or trypsin. Error bars, s.d.; $n = 16$ individual seedlings.



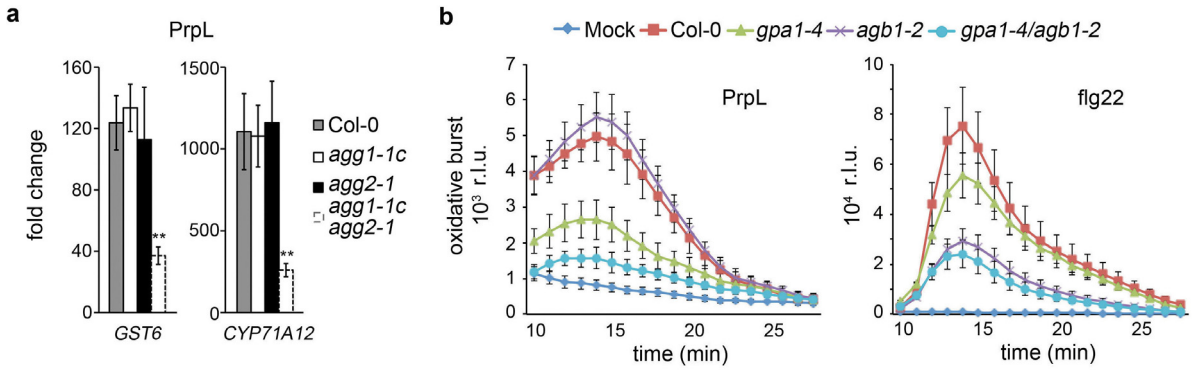
Extended Data Figure 3 | Transcriptional analysis of purified protease IV. **a**, Genome-wide transcriptomic profiles obtained with Affymetrix *Arabidopsis* ATH1 GeneChips of 10-day-old seedlings treated with 20 nM purified PrpL and comparison with published flg22 and oligogalacturonide responses. A Venn diagram shows the similarity of expression behaviour ($|\text{fold change}| > 2$) in response to the three treatments. **b**, Defence gene induction levels measured by RT-qPCR in 10-day-old Col-0 seedlings treated with 20 nM purified PrpL or 20 nM flg22 for 1 h (*WRKY29*, *30*, and *33*) or 6 h (*GST6*, *ERF1*, and *CYP71A12*). Data represent mean ± s.d.; $n = 3$ biological replicates, each containing eight seedlings.



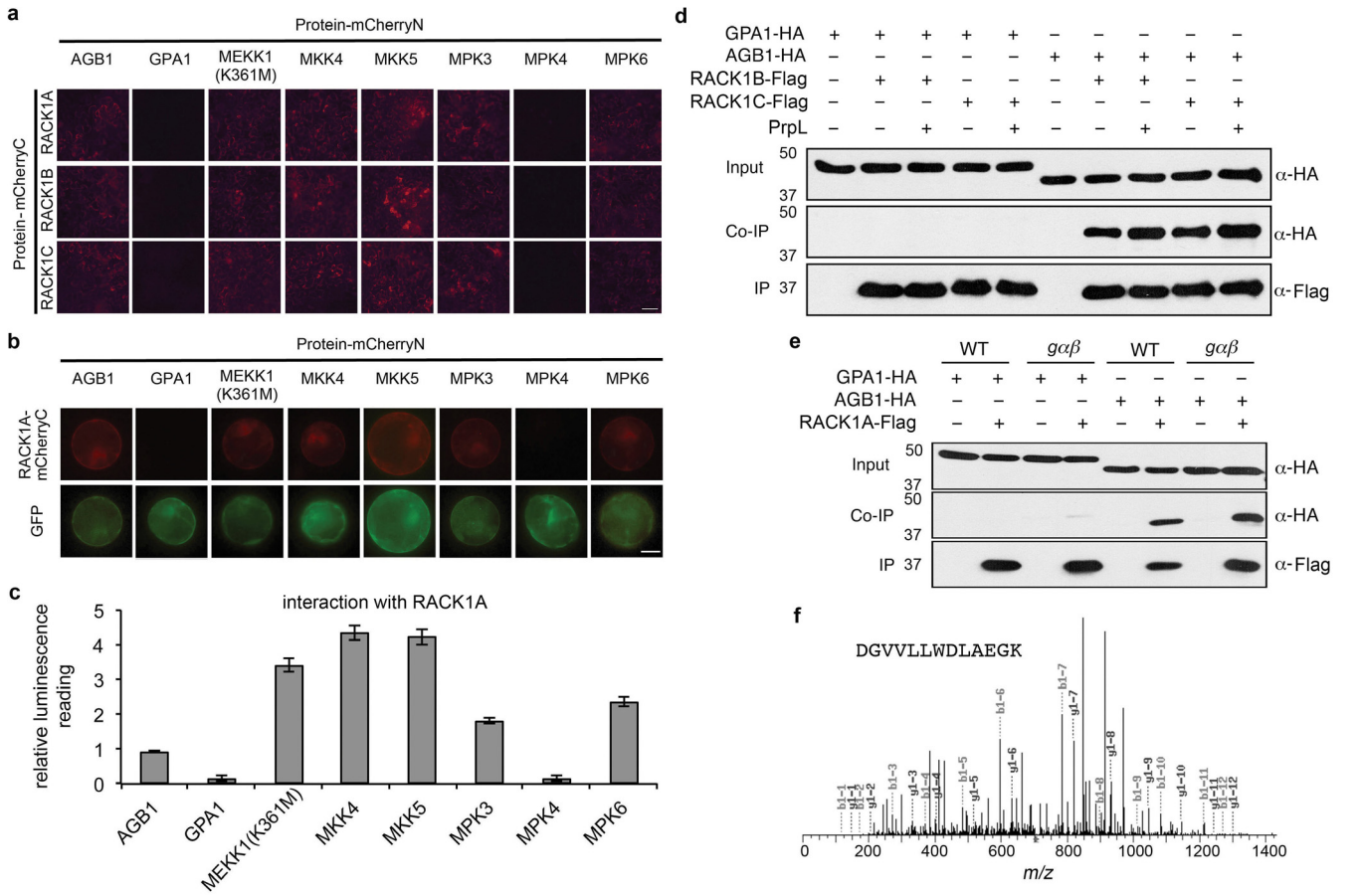
Extended Data Figure 4 | Protease-IV-triggered responses are dependent on proteolytic activity. **a**, Induction of defence-related genes by 20 nM purified PrpL or inactive variants of PrpL measured by RT-qPCR. **b**, Induction of defence-related genes by 20 nM purified PrpL or 20 nM flg22, or 20 nM TLCK-treated PrpL or 20 nM TLCK-treated flg22 measured by RT-qPCR. **c**, Induction of defence-related genes by 20 nM PrpL or 20 nM heat-treated PrpL or 20 nM flg22 or 20 nM heat-treated flg22 measured by RT-qPCR. **d**, Chemiluminescence assay showing elicitation of an oxidative burst by 20 nM purified PrpL, 20 nM inactive variants of PrpL, or 20 nM TLCK-treated PrpL. Data represent mean ± s.d.; $n = 3$ biological replicates with each experiment contains eight seedlings (**a**–**c**) and $n = 16$ individual seedlings (**d**).



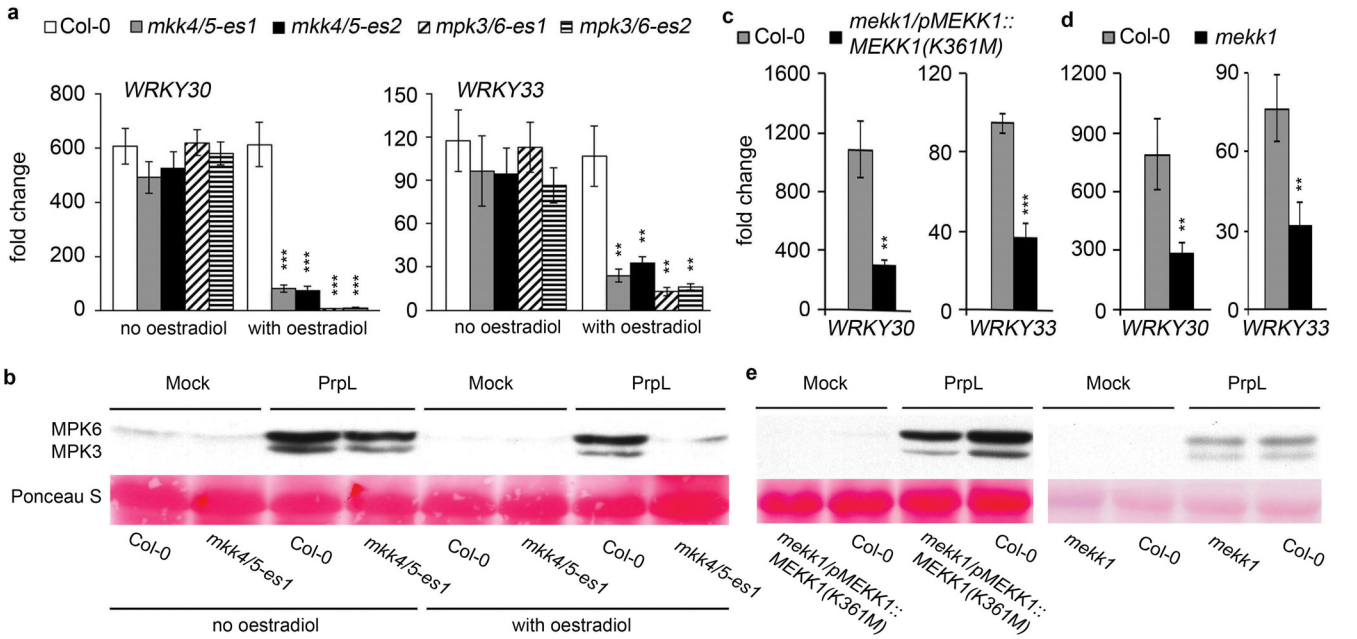
Extended Data Figure 5 | Sequence analyses of *Xanthomonas argC* genes. **a**, The protein sequence alignment between *P. aeruginosa* PA14 PrpL and *X. campestris* pv. *raphani* strain 1946 ArgC (Xcr ArgC). **b**–**d**, Three independent presumptive null mutations in the *Xanthomonas argC* gene: an insertion of G, a single nucleotide mutation, and a deletion. The extra G is highlighted in black in **b**; the single nucleotide substitution is indicated by an arrow in **c**; and the single base deletion is highlighted in black in **d**. The resulting premature stop codons are highlighted in red. Sequences were aligned to the *argC* allele in *X. campestris* pv. *raphani* strain 1946 (Xcr-1946), from which the *argC* gene was cloned. *X. campestris* pv. *campestris* strains 8004 (Xcc-8004); *X. campestris* pv. *campestris* strains BP109 (Xcc-BP109); *X. fuscans* subsp. *fuscans* strain 4834-R (Xf-4834-R); *X. campestris* pv. *vesicatoria* (Xcv). **e**. Activation of *CYP71A12pro:GUS* in 10-day-old seedlings by culture filtrate from *X. campestris* strain Xcr-1946, Xcc-8004, or Xcc-BP109, and *X. campestris* strain 8004 complemented with a functional *argC* gene (8004/*argC*) or transformed with empty vector (8004/vector). Detection of HA-ArgC with an anti-HA antibody. The GUS staining was repeated three times with similar results and the representative images shown were selected from at least three images.



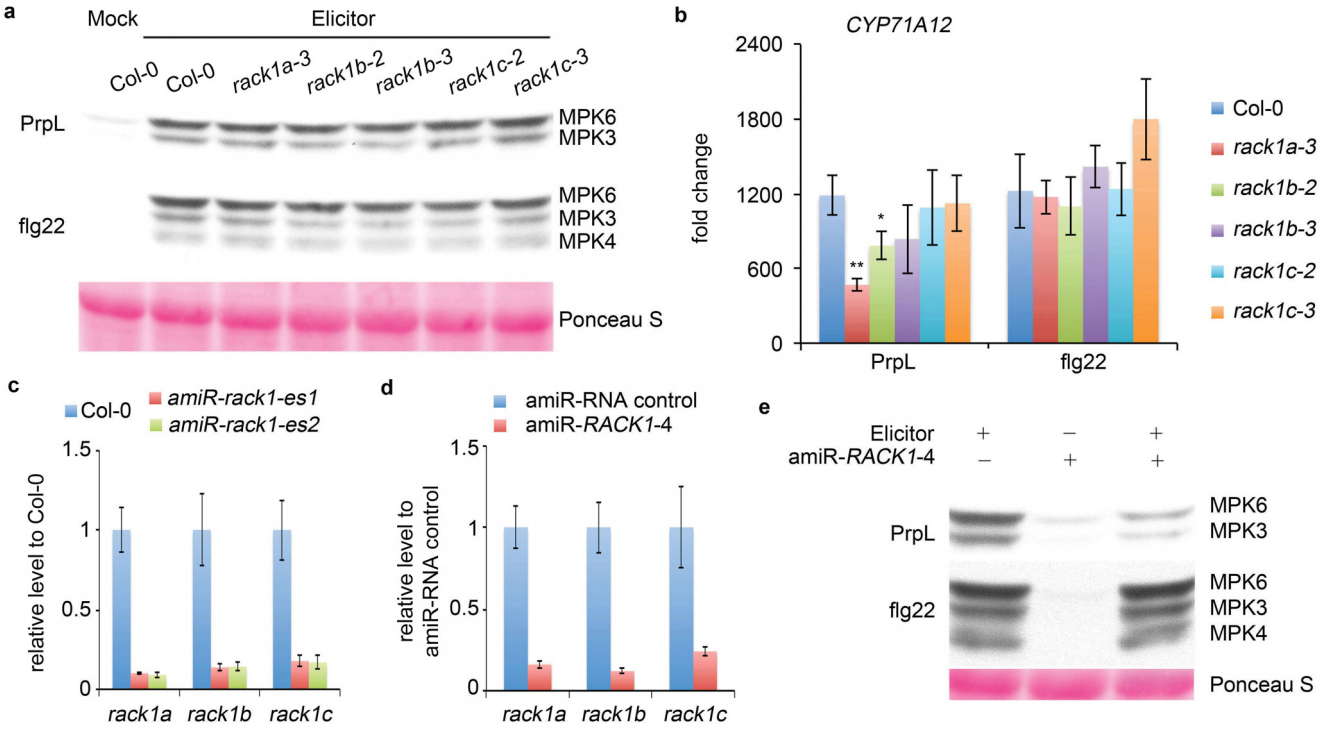
Extended Data Figure 6 | G proteins are required for protease IV response. **a**, Induction of *CYP71A12* and *GST6* gene expression by 20 nM purified PrpL in 10-day-old wild-type Col-0, *gγ* single mutants (*agg1-1c* and *agg2-1*), or a *gγ1γ2* double mutant measured by RT-qPCR. **b**, Chemiluminescence assay showing elicitation of an oxidative burst by 20 nM purified PrpL or 20 nM flg22 in wild-type Col-0 or G-protein tDNA mutants. Data represent mean ± s.d.; $n = 3$ biological replicates with each containing eight seedlings (**a**) and $n = 16$ individual seedlings (**b**); $^{**}P < 0.01$, Student's *t*-test.



Extended Data Figure 7 | Interactions between RACK1 and Gβ or MAPKs. **a**, Split-mCherry assay in 4-week-old *Agrobacterium*-infiltrated *N. benthamiana* leaves. Images were pseudocoloured for visualization. Scale bar, 100 μm. RACK1A, B, C proteins were fused with the C-terminal half of mCherry and the potential interaction partner proteins were fused with the N-terminal half of mCherry. **b**, Split-mCherry assay in *Arabidopsis* protoplasts. RACK1A protein was fused with the C-terminal half of mCherry and the potential interaction partner proteins were fused with the N-terminal half of mCherry. green fluorescent protein (GFP) was included in each experiment to serve as a transfection control. Images were pseudocoloured for visualization. Scale bar, 10 μm. **c**, Relative interaction intensity between RACK1A and G proteins or MAPKs measured by SFLC. RACK1A protein was fused with the FLucN or FLucC to pair with G proteins or MAPKs fused with the other half of firefly luciferase. Both constructs were co-expressed in protoplasts for 6 h and the complemented luciferase activity was used to relatively quantify protein–protein interactions. UBQ10::GUS was included in each experiment to serve as a transfection normalization control. Data represent mean ± s.d.; $n = 3$ technical replicate samples. **d**, Protoplasts were co-transfected with GPA1-HA or AGB1-HA and RACK1B/C-Flag or a control vector. Co-immunoprecipitation was performed with an anti-Flag antibody. Top: the expression of GPA1 or AGB1 protein. Middle: AGB1, but not GPA1, co-immunoprecipitates with RACK1 proteins. Bottom: pulldown of RACK1 proteins by anti-Flag antibody. Protoplasts were treated with 100 nM purified PrpL for 15 min. **e**, Co-immunoprecipitation between GPA1 or AGB1 and RACK1A was performed in wild-type Col-0 or *gαβ* mutant *Arabidopsis* mesophyll protoplasts. Numbers on the left of blots represent marker size in kilodaltons. **f**, Mass spectrophotometric analysis of endogenous proteins pulled down by Flag-tagged MEKK1(K361M). A peptide conserved in all three RACK1 proteins is shown. The experiments in **a** and **b** were repeated three times with similar results.



Extended Data Figure 8 | Protease IV-triggered defence responses in wild-type Col-0 and MAPK mutants. **a**, Induction of *WRKY30* and *WRKY33* gene expression by 20 nM purified PrpL in 7-day-old seedlings of wild-type Col-0 and transgenic *mpk3,6-es1/2* and *mkk4,5-es1/2* plants in the absence or presence of oestradiol. **b**, Western blot depicting activation of MPK3 and MPK6 by 40 nM purified PrpL in 7-day-old seedlings of wild-type Col-0 and transgenic *mkk4,5-es1* plants in the absence or presence of oestradiol. The same molecular mass region of the western blot is shown as in Fig. 1b. **c**, Induction of *WRKY30* and *WRKY33* gene expression by 20 nM purified PrpL in 10-day-old wild-type Col-0 and *mekk1/pMEKK1::MEKK1(K361M)* mutant seedlings. **d**, Induction of *WRKY30* and *WRKY33* gene expression by 20 nM purified PrpL in 4-day-old wild-type Col-0 and *mekk1* null mutant seedlings. **e**, Western blot depicting activation of MPK3 and MPK6 by 40 nM purified PrpL in 10-day-old wild-type Col-0 and *mekk1/pMEKK1::MEKK1(K361M)* mutant seedlings or 4-day-old wild-type Col-0 and *mekk1* null mutant seedlings. The same molecular mass region of the western blot is shown as in Fig. 1b. Data represent mean ± s.d.; $n = 3$ biological replicates with each containing eight seedlings (**a**, **c**, **d**); $^{**}P < 0.01$; $^{***}P < 0.001$, Student's *t*-test versus Col-0 controls.



Extended Data Figure 9 | RACK1 proteins are required for protease IV response. **a**, Western blot depicting activation of MAPKs by 40 nM purified PrpL or 40 nM flg22 in 5-day-old seedlings of wild-type Col-0 and individual *rack1*::tDNA insertion mutants. The same molecular mass region of the western blot is shown as in Fig. 1b. **b**, Induction of *CYP71A12* by 20 nM purified PrpL or 20 nM flg22 in 5-day-old seedlings of wild-type Col-0 and individual *rack1*::tDNA insertion mutants. **c**, RT-qPCR analysis of *rack1a*, *rack1b*, and *rack1c* transcript levels in the 5-day-old Col-0 or *amiR–rack1–es1* and *amiR–rack1–es2* seedlings. **d**, RT-qPCR analysis of *rack1a*, *rack1b*, and *rack1c* transcript levels in *Arabidopsis* protoplasts transfected with amiR–*RACK1*-4 or artificial microRNA control. **e**, Western blot depicting activation of MAPKs by 40 nM purified PrpL or 40 nM flg22 in *Arabidopsis* protoplasts transfected with amiR–*RACK1*-4 or artificial microRNA control. The same molecular mass region of the western blot is shown as in Fig. 1b. Data represent mean ± s.d.; $n = 3$ biological replicates (**b**–**d**); $^{*}P < 0.05$; $^{**}P < 0.01$, Student's *t*-test.

Extended Data Table 1 | *P. aeruginosa* PA14 transposon mutants screened for activation of *CYP71A12pro:GUS*

gene names	gene IDs	type*	gene names	gene IDs	type*	gene names	gene IDs	type*
aprA	865	1	*toxA*	399	2	*clpB*	130	6
aprD	7385	1	*xcpP*	3450	2	*hcp1*	4311	6
aprE	1317	1	*xcpQ*	417	2	*hcpA*	4107	6
aprF	922	1	*xcpR*	812	2	*stnR*	1865	6
aprI	4760	1	*xcpT*	4498	2	*stp1*	3334	6
aprX	1421	1	*xcpW*	3292	2	*vgrG2*	141	6
hasAp	3774	1	*xcpZ*	4249	2	*gacA*	631	R
hasF	1253	1	*xphA*	4239	2	*lasI*	3828	R
cbpD	1394	2	*xqhA*	246	2	*rhlI*	3829	R
cupB5	75	2	*exoT*	7001	3	*rhlR*	3229	R
lasA	1299	2	*exoU*	339	3	*rpoS*	2108	R
lasB	759	2	*exoY*	7430	3	*fimL*	627	S
lipA	2386	2	*pscD*	1303	3	*flgB*	4759	S
lipC	2417	2	*aaaA*	375	5	*flgK*	352	S
pepB	629	2	*eprS*	69	5	*fliC*	1029	S
phoA	914	2	*estA*	354	5	*fliN*	4482	S
phoB	3473	2	*lepA*	631	5	*motA*	2879	S
phoR	1112	2	*tps1*	556	5	*motB*	1935	S
plcB	1956	2	*tps2*	600	5	*pilA*	7353	S
plcH	210	2	*tps3*	555	5	*pilD*	2579	S
plcN	267	2	*tps5*	545	5	*tadZ*	1689	S
pmpA	93	2						

* Numbers represent the type of secretion system. For example, '2' means type II secreted protein or type II secretion machinery protein. R, regulatory proteins; S, surface proteins.

LETTER

doi:10.1038/nature14215

YAP is essential for tissue tension to ensure vertebrate 3D body shape

Sean Porazinski[1]*, Huijia Wang[1]*, Yoichi Asaoka[2]*, Martin Behrndt[3]*, Tatsuo Miyamoto[4]*, Hitoshi Morita[3], Shoji Hata[2], Takashi Sasaki[5], S. F. Gabriel Krens[3], Yumi Osada[6], Satoshi Asaka[2], Akihiro Momoi[6], Sarah Linton[1], Joel B. Miesfeld[7], Brian A. Link[7], Takeshi Senga[8], Atahualpa Castillo-Morales[1], Araxi O. Urrutia[1], Nobuyoshi Shimizu[5], Hideaki Nagase[9], Shinya Matsuura[4], Stefan Bagby[1], Hisato Kondoh[6,10,11], Hiroshi Nishina[2], Carl-Philipp Heisenberg[3] & Makoto Furutani-Seiki[1,6]

Vertebrates have a unique 3D body shape in which correct tissue and organ shape and alignment are essential for function. For example, vision requires the lens to be centred in the eye cup which must in turn be correctly positioned in the head[1]. Tissue morphogenesis depends on force generation, force transmission through the tissue, and response of tissues and extracellular matrix to force[2,3]. Although a century ago D'Arcy Thompson postulated that terrestrial animal body shapes are conditioned by gravity[4], there has been no animal model directly demonstrating how the aforementioned mechano-morphogenetic processes are coordinated to generate a body shape that withstands gravity. Here we report a unique medaka fish (*Oryzias latipes*) mutant, *hirame* (*hir*), which is sensitive to deformation by gravity. *hir* embryos display a markedly flattened body caused by mutation of YAP, a nuclear executor of Hippo signalling that regulates organ size. We show that actomyosin-mediated tissue tension is reduced in *hir* embryos, leading to tissue flattening and tissue misalignment, both of which contribute to body flattening. By analysing YAP function in 3D spheroids of human cells, we identify the Rho GTPase activating protein ARHGAP18 as an effector of YAP in controlling tissue tension. Together, these findings reveal a previously unrecognised function of YAP in regulating tissue shape and alignment required for proper 3D body shape. Understanding this morphogenetic function of YAP could facilitate the use of embryonic stem cells to generate complex organs requiring correct alignment of multiple tissues.

Via exhaustive mutant screening in medaka and zebrafish[5,6], we identified medaka *hir* mutants displaying pronounced body flattening around stage (st.) 25–28 (50–64 h post fertilization, hpf; Fig. 1a). Although general development was not delayed, *hir* mutants exhibited delayed blastopore closure (Fig. 1b, c) and progressive body collapse from mid-neurulation (st. 20, 31 hpf) (Fig. 1d), surviving until just before hatching (6 days post-fertilization, dpf). During body collapse, tissues

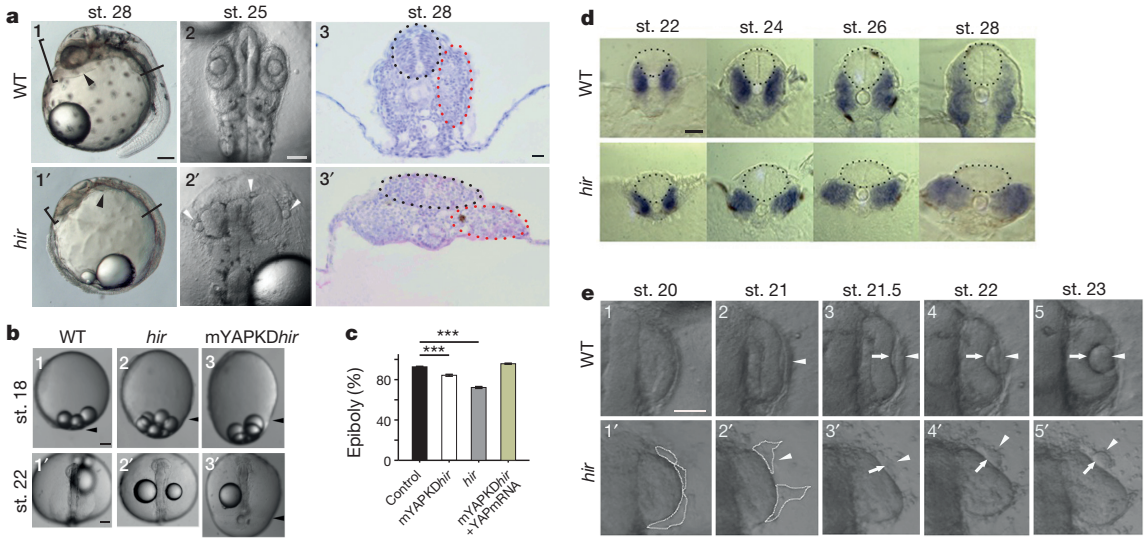


Figure 1 | Organ/tissue collapse and misalignment in *hir* mutants. **a1**, **a1′**, Lateral view of live WT and *hir* mutant embryos, anterior to the left. Arrowheads, heart. Brackets, embryo thickness. **a2**, **a2′**, Dorsal view, anterior upwards. Arrowheads, mislocated lenses. **a3**, **a3′**, Transverse section at the plane shown in **a1** and **a1′**. Neural tubes (black dots) and somites (red dots). **b1**–**b3**, Lateral and **b1′**–**b3′**, dorsal views of live embryos. Arrowheads, blastoderm margin. **c**, Quantification of epiboly (%). Error bars ± s.e.m. (***$P < 0.001$; one-way ANOVA with Dunnett's T3 post hoc. Fig. 1 Source Data). **d**, Transverse sections at 5th somite level, neural tube (encircled) and somites (blue) by *myoD in situ* hybridization. **e**, Time-lapse sequence of dorsal view of WT and *hir* mutant right eyes. Arrowheads, lens placode; arrows, invaginating retina. Fragmented and detaching lens placode demarcated by dotted lines in **e1′** and **e2′**. Scale bars, 40 μm.

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK. [2]Department of Developmental and Regenerative Biology, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan. [3]IST Austria, Am Campus 1, A-3400 Klosterneuburg, Austria. [4]Department of Genetics and Cell Biology, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima 734-8553, Japan. [5]Department of Molecular Biology, School of Medicine, Keio University, Tokyo 160-8582, Japan. [6]Japan Science and Technology Agency (JST), ERATO-SORST Kondoh Differentiation Signaling Project, Kyoto 606-8305, Japan. [7]Department of Cell Biology, Neurobiology, and Anatomy, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. [8]Division of Cancer Biology, Nagoya University Graduate School of Medicine, Nagoya 466-8550, Japan. [9]Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7FY, UK. [10]Graduate School of Frontier Bioscience, Osaka University, Osaka 565-0871, Japan. [11]Faculty of Life Sciences, Kyoto Sangyo University, Kyoto 603-8555, Japan.
*These authors contributed equally to this work.

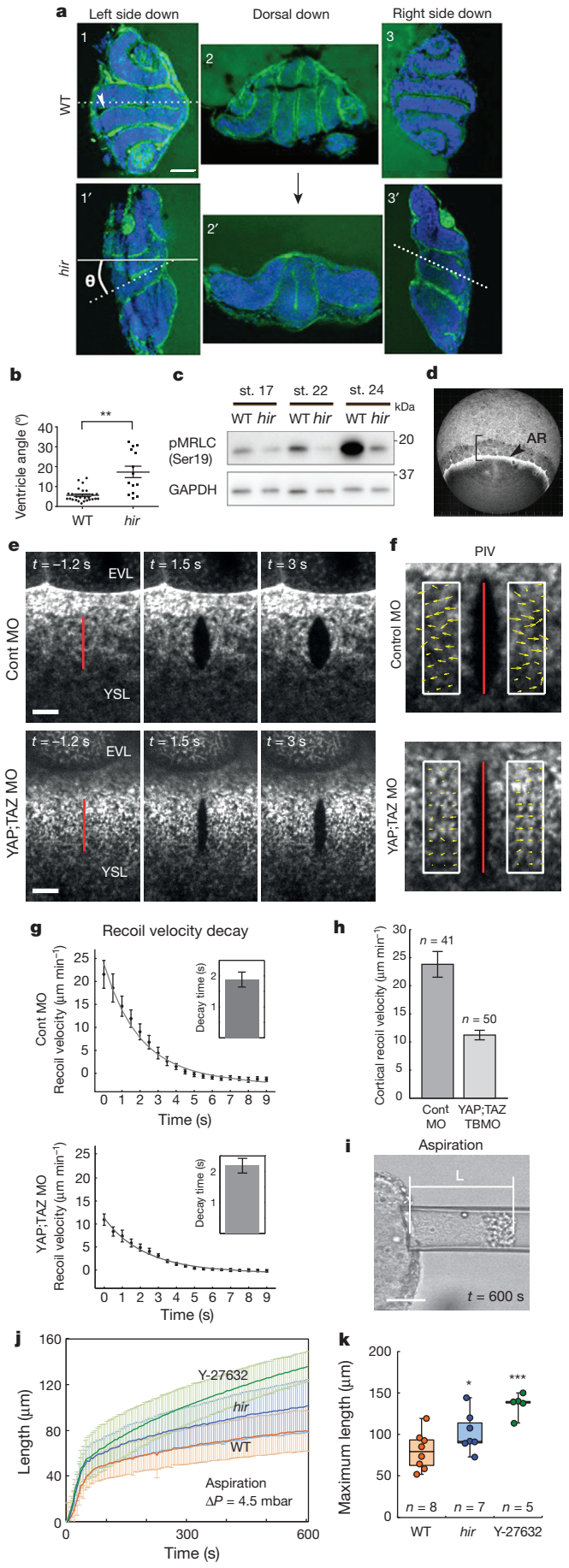


Figure 2 | Tissue tension is reduced in *hir* mutants. **a**, Embryos kept left side down (**a1**, **a1′**), dorsal facing down (**a2**, **a2′**) and right side down (**a3**, **a3′**) from st. 17–25, stained with phalloidin (green, F-actin) and TO-PRO-3 (blue, nucleus). Large black arrow, direction of gravity; θ, angle that the tangent along the brain ventricle (dotted lines in **a1**, **a1′**) makes with horizontal solid line. **b**, Range of collapse of mutant and WT embryos kept sideways. Error bars, ± s.e.m. $^{**}P < 0.01$, *t*-test (see Fig. 2 Source Data). **c**, Immunoblotting of phospho-myosin regulatory light chain (pMRLC, Ser19) and control (GAPDH) (Supplementary Fig. 1). **d**, Actomyosin-labelled *Tg(actb2:myl12.1-eGFP)* zebrafish embryos at 75% epiboly. Arrowhead, YSL actomyosin ring (AR) at the margin of the EVL. Bracket, for analysis of EVL shape anisotropy (Extended Data Fig. 3a). **e**, The actomyosin ring was cut along a 20 μm-long line (red) perpendicular to the EVL/YSL boundary in MO-injected embryos, when control MO injected embryos were at 70–80% epiboly. **f**, Particle image velocimetry (PIV) quantifies the velocity field (yellow arrows) of the recoiling actomyosin network. **g**, Averaged temporal recoil velocity curves, control MO ($n = 41$) and YAP;TAZ KD conditions ($n = 50$). Error bars, error of the mean at 95% confidence. Exponential fit function with a linear offset (black solid line) yields the characteristic decay time (inset) and **h**, the initial recoil velocity for the control MO (23.8 ± 2.3 μm min^{-1}) and YAP;TAZ KD conditions (11.2 ± 0.8 μm min^{-1}). Error bars, 95% confidence interval for the fit results. **i**, Snapshot at the end of aspiration (600 s) of st. 22 neural tube with constant pressure ($\Delta P = 4.5$ mbar). **j**, The curves of the tongue length over time to measure the aspiration of WT, *hir* mutant and ROCK inhibitor (Y27632)-treated neural tube explants. Error bars, ± s.d. Maximum tongue lengths measured at 600 s were compared by *t*-test in **k**. Box plots represent 5%, 25%, median, 75%, and 95%. $^{*}P < 0.05$, $^{***}P < 0.001$. Scale bars, 40 μm in **a**, **i**, 10 μm in **e**.

and organs including neural tube and somites gradually became flattened and improperly aligned (Fig. 1d). Lenses were misaligned outside the eyes (Fig. 1a2, a2′). Mutant lens placodes expressing *sox3* formed normally adjacent to the retina up to st. 20, but then became fragmented and detached from the retina (Fig. 1e1′, e2′, Extended Data Fig. 1a, b and Supplementary Videos 1, 2). These fragments gradually rounded up with some re-attaching to the retina to form ectopic lenses that were not incorporated (Fig. 1e). Thus, tissue flattening and misalignment defects are associated with the flattened mutant phenotype.

Positional cloning identified a nonsense mutation of Leu164 (TTG to TAG) in the WW1 domain of YAP in *hir* (Extended Data Fig. 1c, d). YAP is the nuclear executor of the Hippo pathway and regulates organ growth via stimulation of cell proliferation[7–9]. In wild-type (WT) embryos, YAP transcripts are ubiquitous throughout development[10]. Medaka maternal *YAP* messenger RNA (mRNA) was present at st. 10 in *hir* before onset of zygotic gene expression, but undetectable after st. 18 (Extended Data Fig. 1e). Morpholino oligonucleotide (MO) YAP knockdown (KD) in WT embryos recapitulated the *hir* phenotype (Extended Data Fig. 2a–c, Supplementary Tables 1, 2), and ubiquitous recombinant *YAP* mRNA expression rescued the *hir* phenotype (Extended Data Fig. 1f). In addition, perturbation of maternal *YAP* mRNA translation in *hir* mutant embryos by YAP translation-blocking (TB) MO (mYAP KD *hir* embryos) elicited a more severe blastopore closure and body flattening phenotype than in *hir* zygotic YAP mutants (Fig. 1b3, b3′, c, Supplementary Table 2). Blastopore closure defects, but not flattening, have been reported in YAP KD zebrafish and *Xenopus*[11]. Since TAZ is a functional paralogue of YAP[12], we evaluated its contribution to the YAP KD phenotype in zebrafish. YAP;TAZ double KD zebrafish embryos exhibited more pronounced blastopore closure defects than YAP KD alone (Extended Data Fig. 2d–h). YAP-4SA, which lacks four serines and predominantly localizes to the nucleus[13], rescued the *hir* phenotype more efficiently than WT YAP (Extended Data Fig. 1f), suggesting that the *hir* phenotype depends on nuclear YAP. The main nuclear function of YAP is to promote proliferation and inhibit cell death[14]. *hir* embryos had increased cell death from st. 22 to 26 after body flattening had initiated (increased cell death per se does not lead to body flattening[5,6]). Cell proliferation remained close to normal in *hir* embryos but was strongly suppressed in TAZ KD (and YAP/TAZ double KD) medaka embryos (Extended Data Fig. 2i, j). Thus, in medaka, cell

proliferation is mainly regulated by TAZ, while YAP is predominantly required for 3D body shape.

Three dpf *hir* mutants showed different orientations of body flattening. We therefore examined whether collapse correlated with the direction of gravity. Mutant embryos maintained either right-side or left-side down relative to the earth collapsed towards the earth as indicated by the ventricle tangent (Fig. 2a). Average collapse angle, $\boldsymbol{\theta}$, in mutant embryos was $17.3 \pm 10.7^{\circ}$ ($n = 14$; Fig. 2b) compared to $5.6 \pm 3.3^{\circ}$ ($n = 26, P < 0.01$) in WT. Mutant embryos maintained dorsal side down exhibited apparently uniform dorso-ventral compression (Fig. 2a2, a2′). Thus, flattening in *hir* embryos reflects an inability to withstand external forces (that is, gravity), suggesting reduced tissue tension.

Tissue tension is generated primarily by actomyosin contraction[15]. During WT organogenesis, global levels of phosphorylated myosin regulatory light chain (pMRLC), indicative of actomyosin activity, increased (Fig. 2c), while in *hir* mutants they began decreasing as the blastopore closes (st. 17, 25 hpf), and continued decreasing coincident with tissue collapse and body flattening. To assess tissue tension during blastopore closure, we analysed a surface epithelial cell layer, the enveloping layer (EVL)[16] (Extended Data Fig. 3a1). Comparison of EVL shape anisotropy between WT and *hir* embryos suggested that tissue tension in *hir* is reduced within the EVL (Extended Data Fig. 3a, b). We also quantified actomyosin network tension within the yolk syncytial layer (YSL) of zebrafish embryos with compromised YAP function expressing enhanced green fluorescent protein (EGFP)–myosin light chain protein, Tg(*actb2:myl12.1-eGFP*)[17]. The YSL actomyosin network close to the EVL margin (Fig. 2d) was cut along a 20-μm-long line perpendicular to the margin to reveal circumferential tension (Fig. 2e). Recoil velocities were significantly reduced in YAP;TAZ KD ($n = 50$) compared to control KD embryos ($n = 41$; $11.2 \pm 0.8\ \mu m\ min^{-1}$ vs $23.8 \pm 2.3\ \mu m\ min^{-1}$) (Fig. 2f–h), suggesting reduced actomyosin network tension. Consistent with this, epiboly movements in YAP;TAZ double KD zebrafish embryos were significantly reduced (KD embryos: 53.63 ± 3.93%; control embryos: 70.0 ± 2.18% deep cell epiboly). To test whether reduced actomyosin network tension is also responsible for neural tube tissue flattening in *hir*, we performed micropipette aspiration experiments[18]. *hir* neural explants were significantly less resistant than WT to external forces applied by aspiration, indicating reduced neural tube tissue tension. The higher deformability of *hir* neural tube tissue was paralleled when myosin activity was reduced by ROCK inhibition (Fig. 2i–k). Together, these analyses indicate that YAP is required for actomyosin-mediated tissue tension in medaka and zebrafish.

Single-cell tracking analysis of the growing neural tube in *hir* showed that tissue flattening was associated with failure to stack cells, and increase in cells slipping to one side after perpendicular cell division (Fig. 3a, Extended Data Figs 4, 5). Live imaging showed loss of filopodia that tether lens to retina during lens invagination[1] (Extended Data Figs 1b, 6a, b). The formation of lens–retina filopodia requires fibronectin (FN)-integrin signalling and contractile actomyosin[1]. While st. 22 WT embryos had elongated thin FN fibrils between invaginating lens and retina, *hir* retina showed punctate FN patches (Fig. 3b1′′, b2′′), suggesting defective FN fibril formation. In addition, large ectopic FN deposits were found on the retina in *hir* (Fig. 3b2′). Similar loss of normal FN fibrils and formation of large FN deposits were observed throughout *hir* embryos (Fig. 3b4′, b5′). Furthermore, integrin β1 accumulation between lens and retina was lost in *hir* (Extended Data Fig. 6c). In contrast, cell–cell adhesion and apical markers, including pan-cadherin, atypical PKC (aPKC) and ZO-1, were unaltered in *hir* (data not shown). Mosaic expression of YAP in *hir* and transplantation experiments both showed that the *hir* mutation acts in a non-cell-autonomous manner (Extended Data Fig. 7, Supplementary Table 4). For instance, in invaginated *hir* lens rescued by mosaic expression of YAP, non-YAP expressing *hir* cells recovered filopodia (Extended Data Figs 7b, 6b). These data suggest that YAP functions in tissue alignment by regulating FN assembly.

To identify downstream YAP effectors regulating tissue tension, we used a human 3D spheroid *in vitro* culture system employing the human retina pigmented epithelial cell line hTERT-RPE1 (RPE1), which displayed a relatively mild proliferation defect upon YAP KD. YAP KD spheroids collapsed upon exposure to external forces by slow centrifugation, unlike normal spheroids (Fig. 4a, b). pMRLC levels were reduced in YAP KD spheroids (Fig. 4c), as in *hir*, suggesting that YAP maintains tissue tension also in human 3D tissues. YAP KD spheroids also lacked the typical beehive-like pattern of FN fibrils and, instead, contained large FN deposits, reminiscent of the *hir* retina phenotype

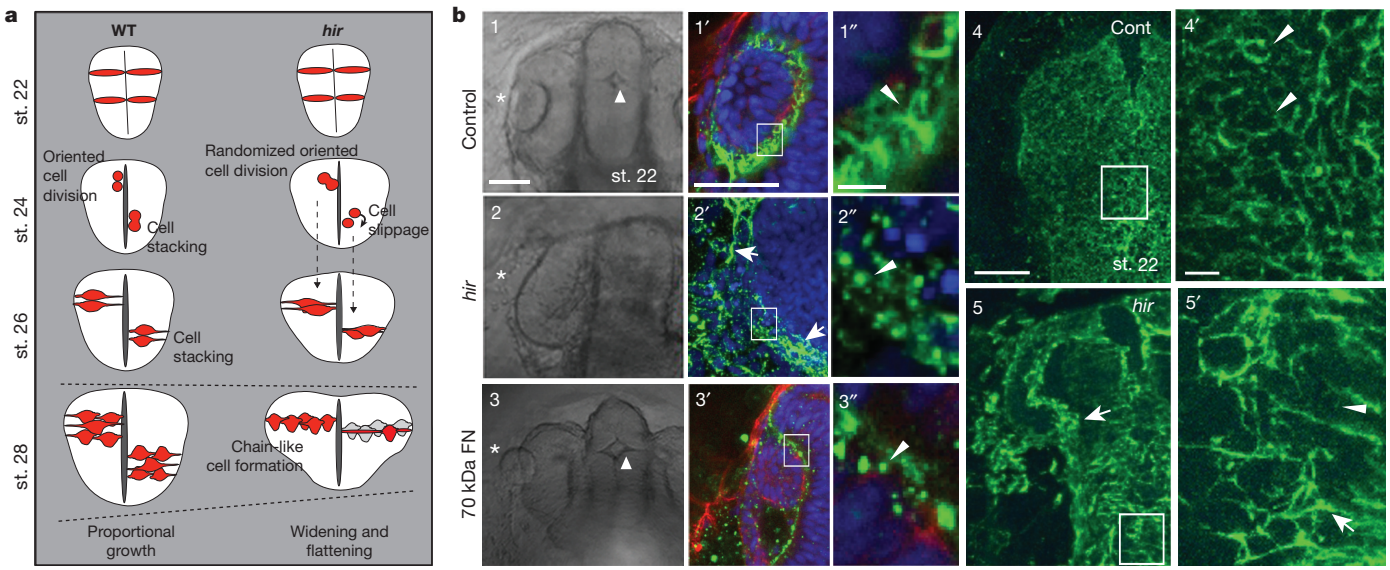


Figure 3 | Cell and tissue dynamics in *hir* mutants. **a**, Schematic: *hir* neural tube collapse is associated with long chain-like arrangements of neuroepithelial cells generated by increased cell slippage and randomized oriented cell division (Extended Data Figs 4, 5). **b**, Whole-mount FN immunohistochemistry (IHC) of st. 22 embryos, dorsal view, anterior to the top. **b1**–**b1′′**, WT embryos injected with out-of-frame 70 kDa N-terminal medaka FN1a+1b mRNA (250 pg) ($n = 20$); **b2**–**b2′′**, uninjected *hir* mutants ($n = 11$); **b3**–**b3′′**, WT embryos injected with N-terminal 70kDa FN1a+1b mRNA (250 pg) ($n = 39$). **b1**–**b3**, Left anterior head of live embryos (asterisks, lens; triangle, forebrain ventricle); **b1′**–**b3′**, left eye of FN IHC (green), boxed area magnified in **b1′′**–**b3′′**; **b4**, **b5**, surface view of FN stained neural tube, WT ($n = 15$) and *hir* ($n = 14$) corresponding to the region in 1 and 2, respectively, boxed area magnified in **b4′** and **b5′**. Arrowheads, FN fibrils/puncta; arrows, FN large deposits. Scale bars, 40 μm in **b1**, **b1′**, **b4**; 5 μm in **b1′′**, **b4′**.

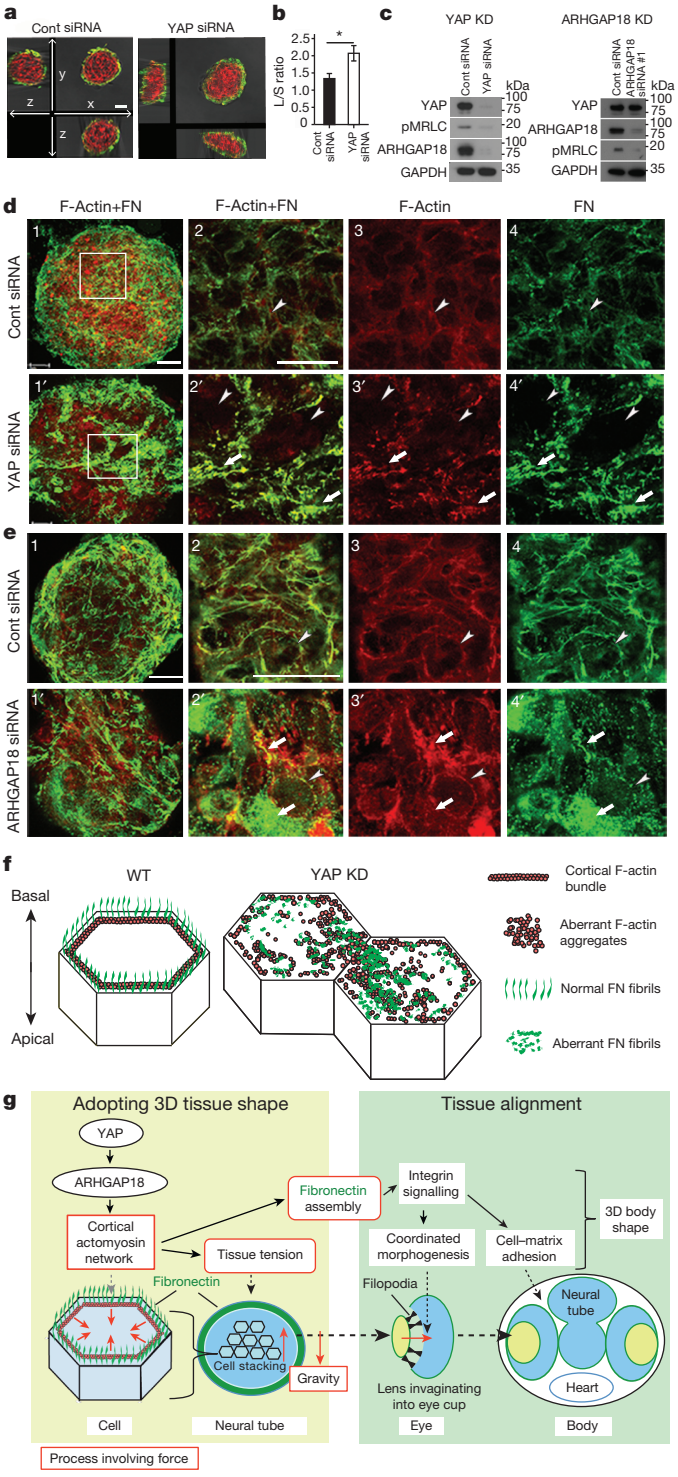


Figure 4 | YAP regulation of tissue tension and FN assembly is mediated by ARHGAP18. **a**, **b**, Confocal 3D sectioning of longest and shortest axes of YAP and control (Cont) KD RPE1 spheroids ($n = 5, 7$) after centrifugation. **b**, Ratio of longest (L)/shortest (S) axes. Error bars, ± s.e.m. $*P < 0.05$, *t*-test (Fig. 4 Source Data). **c**, Immunoblotting of YAP and ARHGAP18 KD spheroids for the indicated proteins (Supplementary Fig. 1). **d**, **e**, Whole-mount imaging of basal surfaces of spheroids transfected with control siRNA ($n = 17$), YAP siRNA ($n = 13$), and ARHGAP18 siRNA ($n = 15$), stained for F-actin (red) and FN (green). **d2**–**d4**, Magnified view of boxed areas in **d1**. Arrowheads, cortical regions; arrows, ectopic F-actin aggregates and aberrant FN fibrils. **f**, Schematic; fine extracellular FN fibrils form in close proximity to cortical F-actin in normal cells, while in YAP and ARHGAP18 KD cells, FN fibrils are reduced and aberrant FN deposits coincide with ectopic F-actin aggregates. **g**, Schematic summarizing how YAP/ARHGAP18-dependent actomyosin network contraction controls tissue shape and alignment. Scale bars, 40 μm in **a**; 30 μm in **d1**, **e1**; 15 μm in **d2**, **e2**.

(Fig. 4d). Cortical actomyosin contraction is required for polymerizing FN monomers to form fibrils[19,20]; consistently, FN fibril formation on the basal surface of control spheroids coincided with cortical F-actin bundles (Fig. 4d). In contrast, loss of normal FN fibrils in YAP KD spheroids was associated with marked reduction of cortical F-actin bundles (Fig. 4d, f). Instead, we observed F-actin aggregates, some of which were associated with large FN deposits, suggesting that they have increased local tension (Fig. 4d). A similar distribution of F-actin and FN was observed in *hir* (Extended Data Fig. 8a). Gene expression profiling of YAP KD spheroids identified only forty genes with reduced expression (see Methods), including *ARHGAP18*, encoding a Rho GTPase activating protein that suppresses F-actin polymerization by inhibiting Rho[21]. *ARHGAP18* transcripts and protein levels were reduced in YAP KD spheroids (Fig. 4c), and ARHGAP18 KD spheroids exhibited a similar phenotype to YAP KD spheroids, including reduced pMRLC levels (Fig. 4c) and aberrant F-actin and FN assembly (Fig. 4e). This suggests that both disruption of cortical F-actin bundles and formation of ectopic F-actin aggregates (Fig. 4f) arise from F-actin over-polymerization in YAP KD spheroids (Extended Data Fig. 8b) and ARHGAP18 KD cells. Together, these results suggest that ARHGAP18 acts downstream of YAP and is required for cortical actomyosin network formation and tissue tension.

To analyse the contribution of actomyosin tension-mediated FN assembly defects to the *hir* eye phenotype, we blocked FN assembly to a similar extent to that in *hir* by overexpressing 70-kDa amino-terminal FN1a and FN1b fragments in WT embryos[22] (Fig. 3b3′, b3′′); this caused near dislocation of the lens and fewer filopodia between lens and retina (Fig. 3b3). *hir* mutants had fewer filopodia than FN assembly blocked embryos (Extended Data Fig. 6a, b), suggesting that contractile actomyosin defects in *hir* exacerbate the incomplete lens dislocation caused by FN assembly defects. In contrast, FN assembly blocked embryos did not exhibit flattened tissues (Fig. 3b1–b3). Furthermore, the medaka FN1 mutant *fukuwarai* (*fku*) also exhibited lens dislocation but not tissue flattening (Extended Data Fig. 8c), suggesting that FN is specifically required for tissue alignment, but not generally for YAP-dependent tissue shape. *ARHGAP18* mRNA levels were significantly reduced in *hir*, and mRNA injection of plasma membrane-targeted myristoylated ARHGAP18 (*myrARHGAP18*) into *hir* substantially rescued FN assembly defects, lens invagination and body flattening (Extended Data Fig. 9a, b). In contrast, inactivation of ARHGAP18 alone was insufficient to produce a recognizable phenotype (data not shown), suggesting that multiple ARHGAP18-related genes function downstream of YAP in medaka embryos. Consistently, short interfering RNA (siRNA) knockdown screening in human cells identified five ARHGAP genes with similar functions to ARHGAP18, homologues of which are conserved in medaka and zebrafish (Extended Data Fig. 9c, d and Supplementary Discussion). These results suggest that ARHGAP18-related genes function as effectors of YAP essential for both tissue shape and FN-dependent tissue alignment. The *hir* phenotype is not simply due to reduced myosin contraction, because injecting mRNA of an activated form of MRLC-DD[23] did not rescue the *hir* phenotype (Extended Data Figs 3a6, b, 8d). Similarly, injection of dominant negative MRLC-AA[23] in WT embryos failed to fully phenocopy the *hir* tissue or body flattening phenotype (Extended Data Fig. 3a5, b). Collectively, these results suggest that YAP function in 3D tissue shape and FN assembly is conserved in human cells and is at least partly mediated by ARHGAP18-related genes.

We propose that YAP is essential for tissue tension, acting through ARHGAP18 and related genes to regulate cortical actomyosin network formation (Fig. 4g). YAP-dependent actomyosin network tension is required for both proper tissue shape and alignment to ensure organ/body shape. Several upstream regulators of YAP-mediated cell proliferation

have been identified, including cellular environment stiffness, suggesting that YAP can function as a mechanosensor[24]. Our data show that YAP also functions as a mechanoregulator of tissue tension. Reduced cortical actomyosin tension is the most probable cause of attenuated tissue tension in *hir* mutants. F-actin over-polymerization perturbs F-actin turnover required for actomyosin contraction in the cytokinetic ring[25]. Our finding that ARHGAP18, a suppressor of F-actin polymerization, functions downstream of YAP further supports a critical role of F-actin polymerization in contractile actomyosin network formation. YAP is required for basal-level actomyosin activity, consistent with ubiquitous expression of actin modulator ARHGAP18[21], additional to which spatiotemporal modulation of actomyosin activity defines tissue shape. Since ARHGAP18 suppresses actin polymerization, which in turn reduces nuclear localization of YAP[26], ARHGAP18 might suppress YAP activity via a negative feedback mechanism. This points to a possible mechanical feedback loop where tissue tension controls YAP, and YAP in turn is required for tissue tension.

Actomyosin contraction promotes FN assembly[27]. The tissue misalignment phenotype in *hir* is most likely owing to failure of YAP-dependent actomyosin contractility in controlling FN assembly. Since FN initiates extracellular matrix organization[27], actomyosin contraction-mediated FN assembly could be a critical *in vivo* mechanism that integrates mechanical signals (for example, tension generated by actomyosin) with biochemical signals (for example, integrin signalling). Notably, the phenotype of YAP KO mouse embryos resembles that of FN KO mouse embryos[28], suggesting that YAP and FN have similar functions in mouse development. Interestingly, while YAP in medaka is predominantly required for tissue tension, its paralogue TAZ seems to be required for cell proliferation (Supplementary Discussion). Given the high degree of conservation of YAP and other Hippo pathway components across metazoa[29], it will be worth investigating whether the extent of tissue three-dimensionality and alignment correlate with the emergence of YAP-mediated resistance to gravity at the evolutionary transition from uni- to multicellular organisms. Finally, generation of a simple organ, such as an eye cup, from induced pluripotent/embryonic stem cells depends on tissue self-organization involving force-mediated processes for which the mechanism remains elusive[30]. Our finding that YAP-dependent force-mediated morphogenesis is required not only for 3D tissue morphogenesis but also tissue alignment suggests that YAP-dependent force-mediated morphogenesis could be involved in self-organization of multiple tissues. Hence, our findings could have implications for the generation of complex organs comprising multiple tissues from induced pluripotent/embryonic stem cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 August; accepted 29 December 2014.
Published online 16 March 2015.

1. Chauhan, B. K. *et al.* Cdc42- and IRSp53-dependent contractile filopodia tether presumptive lens and retina to coordinate epithelial invagination. *Development* **136,** 3657–3667 (2009).
2. Nelson, C. M. & Bissell, M. J. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.* **22,** 287–309 (2006).
3. Mammoto, T. & Ingber, D. E. Mechanical control of tissue and organ development. *Development* **137,** 1407–1420 (2010).
4. Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, 1917).
5. Furutani-Seiki, M. *et al.* Neural degeneration mutants in the zebrafish, *Danio rerio*. *Development* **123,** 229–239 (1996).
6. Furutani-Seiki, M. *et al.* A systematic genome-wide screen for mutations affecting organogenesis in medaka, *Oryzias latipes*. *Mech. Dev.* **121,** 647–658 (2004).
7. Sudol, M. *et al.* Characterization of the mammalian YAP (Yes-associated protein) gene and its role in defining a novel protein module, the WW domain. *J. Biol. Chem.* **270,** 14733–14741 (1995).
8. Pan, D. The Hippo signaling pathway in development and cancer. *Dev. Cell* **19,** 491–505 (2010).
9. Zhao, B., Tumaneng, K. & Guan, K.-L. L. The Hippo pathway in organ size control, tissue regeneration and stem cell self-renewal. *Nature Cell Biol.* **13,** 877–883 (2011).
10. Miesfeld, J. B. & Link, B. A. Establishment of transgenic lines to monitor and manipulate Yap/Taz-Tead activity in zebrafish reveals both evolutionarily conserved and divergent functions of the Hippo pathway. *Mech. Dev.* **133,** 177–188 (2014).
11. Gee, S. T., Milgram, S. L., Kramer, K. L., Conlon, F. L. & Moody, S. A. Yes-associated protein 65 (YAP) expands neural progenitors and regulates Pax3 expression in the neural plate border zone. *PLoS ONE* **6,** e20309 (2011).
12. Lei, Q. Y. *et al.* TAZ promotes cell proliferation and epithelial-mesenchymal transition and is inhibited by the Hippo pathway. *Mol. Cell. Biol.* **28,** 2426–2436 (2008).
13. Zhao, B., Li, L., Tumaneng, K., Wang, C. Y. & Guan, K.-L. L. A coordinated phosphorylation by Lats and CK1 regulates YAP stability through $SCF^{\beta\text{-}TRCP}$. *Genes Dev.* **24,** 72–85 (2010).
14. Heisenberg, C.-P. P. & Bellaïche, Y. Forces in tissue morphogenesis and patterning. *Cell* **153,** 948–962 (2013).
15. Vicente-Manzanares, M., Ma, X., Adelstein, R. S. & Horwitz, A. R. Cytoskeletal motors: non-muscle myosin II takes centre stage in cell adhesion and migration. *Nature Rev. Mol. Cell Biol.* **10,** 778–790 (2009).
16. Köppen, M., Fernández, B. G., Carvalho, L., Jacinto, A. & Heisenberg, C.-P. P. Coordinated cell-shape changes control epithelial movement in zebrafish and *Drosophila*. *Development* **133,** 2671–2681 (2006).
17. Behrndt, M. *et al.* Forces driving epithelial spreading in zebrafish gastrulation. *Science* **338,** 257–260 (2012).
18. Guevorkian, K., Colbert, M.-J., Durth, M., Dufour, S. & Brochard-Wyart, F. Aspiration of biological viscoelastic drops. *Phys. Rev. Lett.* **104,** 218101 (2010).
19. Singh, P., Carraher, C. & Schwarzbauer, J. E. Assembly of fibronectin extracellular matrix. *Annu. Rev. Cell Dev. Biol.* **26,** 397–419 (2010).
20. Rolo, A., Skoglund, P. & Keller, R. E. Morphogenetic movements driving neural tube closure in *Xenopus* require myosin IIB. *Dev. Biol.* **327,** 327–338 (2009).
21. Maeda, M. *et al.* ARHGAP18, a GTPase-activating protein for RhoA, controls cell shape, spreading, and motility. *Mol. Biol. Cell* **22,** 3840–3852 (2011).
22. McDonald, J. A. *et al.* Fibronectin's cell-adhesive domain and an amino-terminal matrix assembly domain participate in its assembly into fibroblast pericellular matrix. *J. Biol. Chem.* **262,** 2957–2967 (1987).
23. Iwasaki, T., Murata-Hori, M., Ishitobi, S. & Hosoya, H. Diphosphorylated MRLC is required for organization of stress fibers in interphase cells and the contractile ring in dividing cells. *Cell Struct. Funct.* **26,** 677–683 (2001).
24. Dupont, S. *et al.* Role of YAP/TAZ in mechanotransduction. *Nature* **474,** 179–183 (2011).
25. Pinto, I. M., Rubinstein, B., Kucharavy, A., Unruh, J. R. & Li, R. Actin depolymerization drives actomyosin ring contraction during budding yeast cytokinesis. *Dev. Cell* **22,** 1247–1260 (2012).
26. Sansores-Garcia, L. *et al.* Modulating F-actin organization induces organ growth by affecting the Hippo pathway. *EMBO J.* **30,** 2325–2335 (2011).
27. Daley, W. P., Peters, S. B. & Larsen, M. Extracellular matrix dynamics in development and regenerative medicine. *J. Cell Sci.* **121,** 255–264 (2008).
28. Morin-Kensicki, E. M. *et al.* Defects in yolk sac vasculogenesis, chorioallantoic fusion, and embryonic axis elongation in mice with targeted disruption of Yap65. *Mol. Cell. Biol.* **26,** 77–87 (2006).
29. Hilman, D. & Gat, U. The evolutionary history of YAP and the Hippo/YAP pathway. *Mol. Biol. Evol.* **28,** 2403–2417 (2011).
30. Sasai, Y. Cytosystems dynamics in self-organization of tissue architecture. *Nature* **493,** 318–326 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Raff, T. Perry, A. Ward, M. Wills, J. Caunt, J. Clarke, L. Hurst and C. Tickle for critical reading and comments. We thank M. Tada, M. Furuse, N. Wada, Y. Nakai, J. Robinson and R. Kelsh for contributions to the paper and University of Bath for fish and bioimaging facilities. This work was funded by the ERATO/SORST projects of JST, Japan (H.K.), National Institutes of Health R01EY014167 (B.A.L.) and Medical Research Council, UK (M.F.-S.).

Author Contributions S.P., H.W., Y.A., M.B., T.M., H.M., S.H., T.S., S.F.G.K., Y.O., S.A., A.M., S.L., J.B.M., B.A.L., T.S., A.C.M., A.O.U., S.B. and M.F.-S. performed experiments. S.P., H.W., Y.A., M.B., T.M. and M.F.-S. conceived the study. S.B., N.S., H.N., S.M., H.K., C.-P.H., H.N. and M.F.- S. supervised the study. C.-P.H. and M.F.- S. wrote the paper. All authors interpreted data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.-P.H. (heisenberg@ist.ac.at), H.N. (nishina.dbio@mri.tmd.ac.jp) or M.F.-S. (furutaniseiki@gmail.com).

METHODS

Fish maintenance and fish strains. Medaka (*Oryzias latipes*) and zebrafish (*Danio rerio*) strains were maintained and raised according to previously published procedures[31]. Medaka and zebrafish embryos were raised in E3 solution at 28 °C. Fish care and procedures were approved by the University of Bath Ethical Review Committee, and are in compliance with the Animals Scientific Procedures Act 1986 of the UK. Medaka WT strains K-Cab, K-Kaga, and the mutant strain *hir*$^{54\text{-}20C}$, *fku*$^{8\text{-}33A}$, were used[6]. Zebrafish WT strain AB and *Tg(actb2:myl12.1-eGFP)*[32], that allow visualization of myosin, were used.

Embryological experiments. For fixation and live imaging, embryos were anaesthetized with 0.01% tricaine. For live imaging, embryos were embedded in 0.8% low melting temperature agarose (Type IV-A, Sigma, USA) in 35 mm glass-base dishes (Iwaki, Japan) at 28 °C. Standard embryological procedures including, dechorionation, fixation, *in situ* hybridization, immunohistochemistry, microinjection and cell transplantation were carried out according to previously published procedures[31]. Cells were transplanted to the region fated to become the eye and Cuvier's duct according to our fate map[33].

Positional cloning of *hir*. The *hir* mutation induced in the K-Cab strain was crossed with the polymorphic K-Kaga strain to carry out genetic mapping according to a previously published procedure[34]. To map the *hir* mutation on the chromosome, bulked segregant analysis was performed using M-markers[35] on DNA isolated from 48 homozygous mutant embryos and 48 WT siblings from F_2 embryos of mutant × K-Kaga mapping crosses.

Chromosome walking on chromosome 13 was performed using restriction fragment length polymorphism markers between K-Cab and K-Kaga strains to map to the two BAC clones. For fine mapping, 1,908 meioses were analysed to identify 9 recombinants mapping *hir* mutation close to *YAP*. *YAP* complementary DNA was amplified from *hir* mutants by RT–PCR and sequenced directly to identify the mutation.

RT–PCR cDNA cloning and construction. Total RNAs were isolated using TRIzol (Life Technologies) and were converted to cDNA using the RNA-PCR kit ver.3 (Takara Bio, Japan) followed by PCR using KOD plus polymerase (Toyobo, Japan). For mRNA production, PCR amplified full-length cDNAs (medaka *YAP, 70KDaFN1a,b, ARHGAP18*) were cloned into pCS2+ and for *in situ* hybridization medaka *sox3* cDNA was cloned into pBluescript II SK(−). *pCS2+myr-ARHGAP18* was constructed by adding the myristoylation sequence using oligonucleotides to produce myristoylated *ARHGAP18* mRNA. mRNAs were synthesized using SP6 mMESSAGE mMACHINE Kit (Ambion, USA). Primer sequences are shown in Supplementary Table 5.

Gravity experiment. Dechorionated embryos were embedded in 0.8% low melting temperature agarose in three orientations against gravity at st. 17, fixed at st. 25 and subjected to cryosectioning to determine the direction of tissue/organ collapse. Collapse of embryos towards gravity was assessed using images of sections stained with TO-PRO-3 and phalloidin.

Microinjection. mRNA, DNA and Morpholino were injected at 1-cell or 8-cell stages to deliver them to all cells or in a mosaic manner. The volume of one-shot of injection was 0.5 nl.

Phenotypic rescue experiments. Embryos from *hir*$^{+/-}$ heterozygote crosses were injected with mRNA of YAP variants. For transplantation phenotypic rescue experiments, embryos were genotyped by PCR using primers (Supplementary Table 5).

Morpholino KD analysis in medaka and zebrafish. Morpholino oligonucleotide (MO) from Gene Tools (USA) were used (Supplementary Table 6). Specificity of KD by MO was confirmed in a slightly different manner in medaka and zebrafish. Since rescue of the phenotype by mRNA injection did not work effectively in zebrafish, three different types of MOs, translation blocking (TB), splicing blocking (SB) and 5′ UTR MOs, were used and all were confirmed to induce a similar phenotype. In medaka, TB and SB MOs were used, and the phenotype was rescued by co-injecting corresponding mRNAs. To determine efficiency of KD, semiquantitative RT–PCR was carried out using primers that distinguish defective splicing from normal forms of mRNA (Supplementary Table 5).

Immunohistochemistry. Embryos were fixed in either 4% paraformaldehyde (PFA), Dent fixative or 1% TCA for 1–3 days and subjected to cryosectioning as described previously[31]. Antibodies used were: anti-FN antibody (Ab), Sigma F3648 at 1:100; β-integrin monoclonal Ab, 8c8 (Developmental Studies Hybridoma Bank, USA) at 1:10; anti-aPKC C-20 (SC216, Santa Cruz Biotech, USA) at 1:100; anti-PCNA (PC10, Santa Cruz Biotech, USA) at 1:500; anti-laminin (Ab-1, NeoMarkers, USA) at 1:100 and anti-ZO-1[36] (gift from M. Itoh) at 1:1. Sections were counterstained with Alexa Fluor 488 or 546 Phalloidin (A12379, A22283, Invitrogen USA) at 1:250 and TO-PRO-3 (T3605, Invitrogen, USA) at 1:1,000.

Time-lapse microscopy and image analysis. Time-lapse analysis of lens dislocation was carried out using a Leica MZ16FA dissecting microscope. Confocal microscopy used a Leica TCS SP5 and images were analysed by Imaris 7.3 (Bitplane, ANDOR Technology, UK) and Amira 5.1 (Visage Imaging, USA). Cell division orientation (θ) of telophase cells in time-lapse sequences was determined by drawing an axis from the ventricular zone-attached non-moving daughter cell (asterisk Extended Data Fig. 5c) towards the non-attached moving daughter cell[37]. The acute angle of this axis was then measured against the axis of the ventricular zone. Imaging was carried out dorsal side down using an inverted microscope. Rose diagrams were generated using Oriana v4 (Kovach Computing Services, UK).

Spheroid analysis. hTERT-RPE1 cells (American Type Culture Collection; CRL-4000) were seeded (2×10^5 cells per well in 6-well plates). Each stealth RNA (100 pmol) of Opti-Mem medium (Life Technologies) was transfected using Lipofectamine RNAi Max (Life Technologies) followed by incubation for 24 h at 37 °C. Trypsin treatment was used to collect RNAi-transfected cells from wells which were resuspended in 2 ml of 10% FBS (Hyclone, ThermoFisher Scientific)-DMEM. These resuspensions were seeded to 6 wells of a 12-well plate (Hydrocell, CellSeed Japan) and incubated for 48 h at 37 °C. Spheroids were fixed in 3% formalin and subjected to immunostaining. Reagents used for immunostaining: anti-β-catenin (BD transduction, 610154, 1:200), anti-FN (Sigma F3648, 1:500), Alexa Fluor 546 Phalloidin (Invitrogen, A22283,1:200). For the list of primers see Supplementary Table 5.

Western blotting. Spheroids were lysed in lysis buffer (0.5% TritonX-100, 150 mM NaCl, 20 mM Tris-HCl pH 7.5). The lysates were sheared with a 21-gauge needle, incubated on ice for 30 min and clarified by centrifugation at 20,817*g* for 15 min at 4 °C. The extracted proteins were separated by SDS–PAGE and transferred to immobilon transfer membrane (Millipore) for western blotting analyses. The primary antibodies were anti-YAP1 pAb (4912 Cell Signaling, 1:500), anti-fibronectin pAb (F3648, Sigma Aldrich, 1:1,000), anti-ARHGAP18 pAb (1:10,000)[17,21], anti-MYH9 pAb (3403 Cell Signaling, 1:1,000), anti-Phospho Ser1943-MYH9 pAb (5026 Cell Signaling, 1:1,000), anti-MYH10 mAb (8824 Cell Signaling, 1:1,000), anti-Phospho-Ser19 MLC2 (3675, Cell Signaling, 1:100), and anti-GAPDH mAb (sc32233, Santa Cruz, 1:5,000).

Actomyosin tension measurement by laser cutting. Laser cutting experiments were carried out using a UV-laser ablation system as previously described[17]. Tg(*actb2:myl12.1-eGFP*)[32] embryos were mounted in 1% low melting point agarose (Invitrogen) embedded in E3 medium inside a glass bottom Petri dish (Mattek). A 63× water immersion objective (NA = 1.2, Zeiss) was used to visualize the YSL actomyosin ring at respective epiboly stages. Cuts were made at a distance of 20 μm from the EVL/YSL boundary by applying 25 UV pulses at 1 kHz to 40 equidistant sites along a 20-μm-long line perpendicular to the EVL margin as depicted in Fig. 2e. Fluorescent images of embryos were captured using an iXon DU-897-BV camera (Andor Technology) with a 380 ms exposure time and 500 ms frame rate (LabVIEW v10.0.1). The ablation procedure itself took 1.2 s during which no images were acquired. Temperature was kept constant at 28.5 ± 1 °C throughout the experiment by means of a custom-built temperature chamber and an objective heating ring. The recoil velocity of the cortex in response to the cut opening was analysed using customized Matlab (v7.12) scripts based on particle image velocimetry (PIV) as previously described[17,38]. The component of the PIV flow field that is orthogonal to the cut line was averaged in two adjacent rectangles (Fig. 2f) for time frames up to 9 s post-ablation. The resulting recoil velocity curves for single embryo ablation experiments were averaged to yield the mean temporal recoil velocity curve for the depicted conditions (Fig. 2g). Laser ablation experiments that caused wound response recognizable by a strong accumulation of myosin following the ablation were discarded from the analysis. In these experiments leakage of yolk cytoplasm through a membrane opening may interfere with the cortical tension measurements[17].

Micropipette aspiration analysis. The whole neural tube was dissected out from st. 22 medaka embryos and was cut using a tungsten needle at the level of diencephalon-midbrain boundary. The micropipette was connected to a Microfluidic Flow Control System (Fluigent, Fluiwell) which was controlled via a custom-programmed Labview (National Instruments) interface. In the BSS medium, the neural tube was aspirated from the open end by a micropipette (internal radius = 30–35 μm) at a constant pressure ($\Delta P = 4.5$ mbar) for 10 min. Aspiration was imaged at 500 ms intervals by a Leica SP5 inverted confocal microscope using a Leica 20×, 0.7 NA objective. Temperature in the dish was kept constant at 28 °C by a heated sample holder. Measuring the tongue length of the tissue within the micropipette using FIJI software over time yielded the characteristic tissue flow curves during aspiration for WT and *hir* mutant neural tube explants. To reduce cortical tension WT neural tube explants were treated with ROCK inhibitor Y27632 (250 μM dissolved in water) for 15 min before performing the micropipette aspiration experiment.

Oligo DNA microarray analysis. For the Oligo DNA microarray analysis, total RNA samples were collected from hTERT-RPE1 multicellular spheroids. 3D-Gene Human Oligo chip 25k (TORAY) was used. Total RNA of YAP siRNA-transfected spheroids and that of negative control siRNA were labelled with Cy3 or Cy5 using the Amino Allyl MessageAMP II aRNA Amplification Kit (Life Technologies), respectively. The Cy3- or Cy5-labelled aRNA pools and hybridization buffer were mixed, and hybridized for 16 h at 37 °C. The hybridization was performed using

the supplier's protocols (http://www.3d-gene.com). Hybridization signals were scanned using a 3D-Gene Scanner 3000 (TORAY). Detected signals for each gene were normalized by a global normalization method (Cy3/Cy5 ratio median = 1). Genes with Cy3/Cy5 normalized ratios greater than 2.0 or less than 0.5 were defined, respectively, as commonly up- or downregulated genes. The results were deposited at GEO under the accession number GSE54146.

Quantitative RT–PCR analysis. Total RNA was isolated from WT and *hir* mutant embryos at various developmental stages using TRIzol (Invitrogen) according to the manufacturer's instructions. First-strand cDNA was synthesized from 1 μg total RNA using Superscript III reverse transcriptase (Invitrogen) with an oligo-dT primer. Each quantitative real-time RT–PCR was performed using the CFX96 real-time PCR detection system (Bio-Rad). Primers used for RT–PCR analysis are shown in Supplementary Table 5. For a 10 μl PCR, cDNA template was mixed with the primers to final concentrations of 250 nM and 5 μl of SsoFast EvaGreen Supermix (Bio-Rad), respectively. The reaction was first incubated at 95 °C for 3.5 min, followed by 45 cycles at 95 °C for 30 s, 65 °C for 30 s and 72 °C for 30 s.

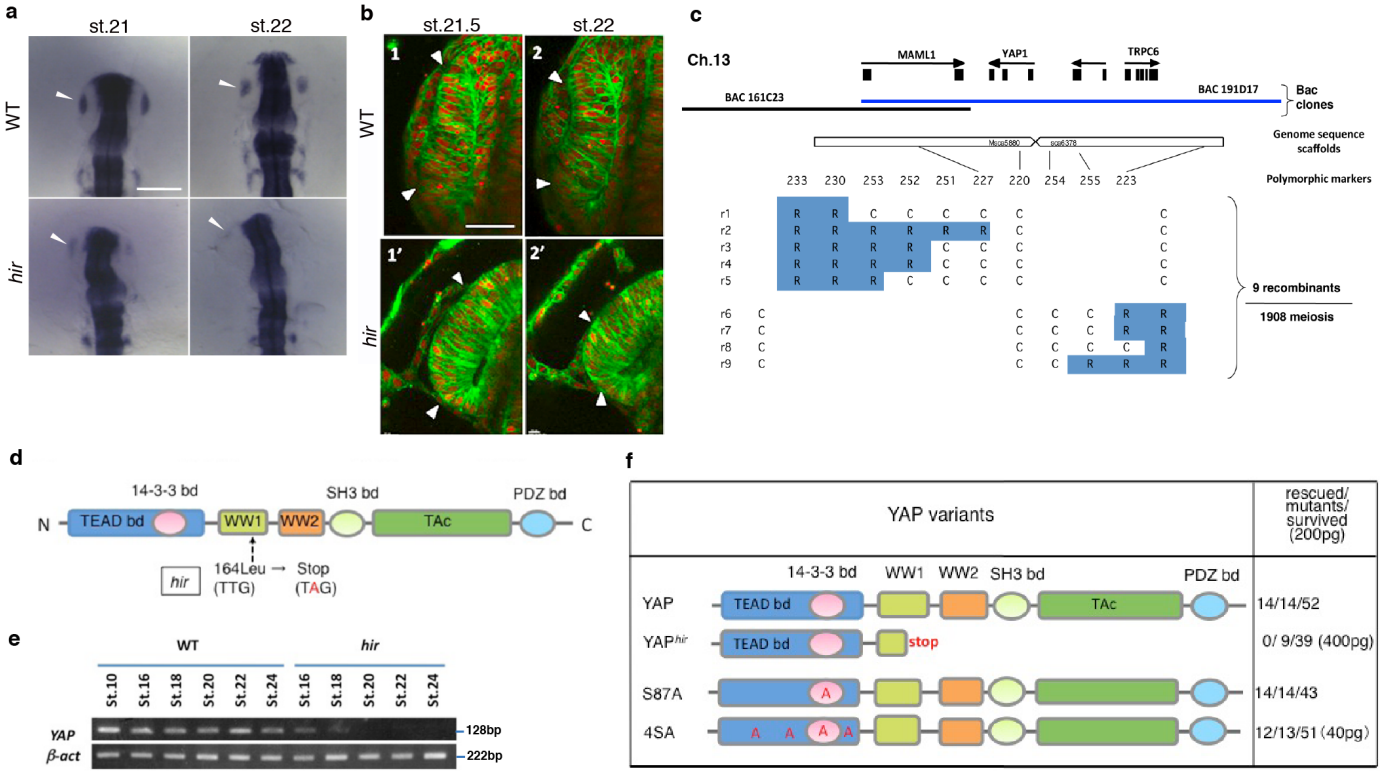
Phylogenetic analysis of ARHGAP18-related genes in 11 metazoan species. Lists of homologues of ARHGAP18 family (TF314044) and its closely related families ARHGAP6 (TF316710) and ARHGAP11 (TF332212) in 11 metazoan model species were downloaded from the Treefam database. Amino-acid sequences for these genes were downloaded from Ensembl. Multiple sequence alignment was performed using the PRANK package. This alignment was used to infer the phylogenetic relationship of these genes using maximum likelihood using FastTree 2.1.

ARHGAP siRNA screening in HeLa cell line. A library of siRNAs targeting human GAPs was obtained from Invitrogen. HeLa cells cultured in 24-well plates were transfected with siRNAs (20 nM) using Lipofectamine RNAiMAX. After 72 h, cells were fixed with 4% PFA and stained with FITC-labelled phalloidin (Invitrogen). Images were taken using an Olympus IX71 fluorescence microscope.

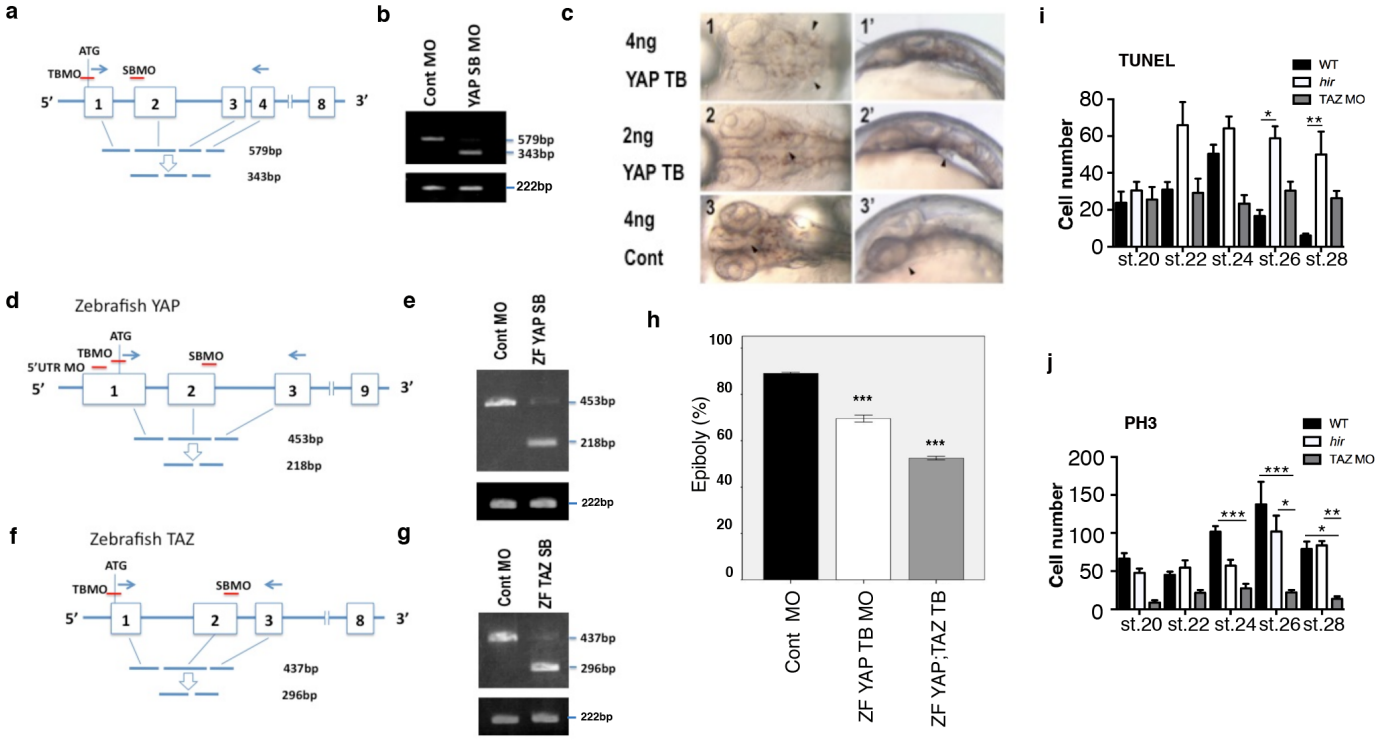
Statistical analyses. Statistical significance between WT and mutant groups was tested using independent two-tailed *t*-tests (for two-way comparisons) and one-way ANOVAs (for multiple comparisons), with a Dunnett's T3 post-hoc where necessary, in SPSS 20 (IBM) or Prism v5.0 (GraphPad). The Dunnett's T3 post-hoc assumes variances to be unequal and allows comparisons of groups with different *n* numbers. To test for differences in mitotic orientation between WT and *hir* we performed the Kolmogorov–Smirnov (KS) test (http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html). The KS test makes no assumptions about the distribution of data being tested. Sample size was not pre-determined. We repeated experiments a minimum of three times with sufficient *n* numbers for each repeat to be confident that reported results are representative. Randomization was not applied to allocate embryos to experimental groups. Blinding to group allocation was not used. Error bars on graphs show ± standard error of the means (s.e.m.), except when stated otherwise. Data points that deviated by more than ± 3× the standard deviation of the sample mean were excluded from analysis.

***P* values and sample sizes.** *P*-values vs WT unless specified. Fig. 1c: $n_{cont} = 39$, $n_{hir} = 25$ ($P = < 0.001$), $n_{mYAPKDhir} = 24$ ($P = < 0.001$), $n_{mYAPKDhir+YAPmRNA} = 22$ ($P = 1.000$). Fig. 2b: $n_{WT} = 26$, $n_{hir} = 14$ ($P = 0.0001$). Fig. 4b: $n_{contsiRNA} = 7$, $n_{YAPsiRNA} = 5$ ($P = 0.023$). Extended Data Fig. 2h: $n_{contMO} = 20$, $n_{ZFYAPTBMO} = 11$ ($P = < 0.001$), $n_{ZFYAP;TAZTB} = 10$ ($P = < 0.001$). Extended Data Fig. 2i: n_{WT} st. 20 = 5, st. 22 = 8, st. 24 = 13, st. 26 = 6, st. 28 = 10, n_{hir} st. 20 = 8, st. 22 = 4, st. 24 = 6, st. 26 = 9 ($P = 0.0284$), st. 28 = 5 ($P = 0.0088$), n_{TAZMO} st. 20 = 5, st. 22 = 5, st. 24 = 10, st. 26 = 12, st. 28 = 12. Extended Data Fig. 2j: n_{WT} st. 20 = 11, st. 22 = 7, st. 24 = 10, st. 26 = 11, st. 28 = 11, n_{hir} st. 20 = 7, st. 22 = 7, st. 24 = 11, st. 26 = 13 ($P = 0.0158$ vs TAZMO st. 26), st. 28 = 7 ($P = 0.0075$ vs TAZMO st. 28), n_{TAZMO} st. 20 = 5, st. 22 = 5, st. 24 = 10 ($P = 0.0007$), st. 26 = 8 ($P = 0.0008$), st. 28 = 6 ($P = 0.0120$). Extended Data Fig. 3b: $n_{WT} = 174$, $n_{hir} = 70$ ($P = < 0.001$), $n_{mYAPKDhir} = 85$ ($P = < 0.001$), $n_{MRLCAA > WT} = 135$ ($P = < 0.001$), $n_{MRLCDD>hir} = 92$ ($P = 0.1830$ vs *hir*). Extended Data Fig. 4b: $n_{WT} = 3$, $n_{hir} = 3$. Extended Data Fig. 5b: n_{WT} cell stacking = 9, cell slippage = 8, parallel division = 5, n_{hir} cell stacking = 3 ($P = <0.01$), cell slippage = 21 ($P = <0.05$), parallel division = 5. Extended Data Fig. 5d: KS-test, $P = 0.01$, n_{WT} st. 22–24 = 32, st. 25–26 = 13, n_{hir} st. 22–24 = 14, st. 25–26 = 20. Extended Data Fig. 6b: $n_{WT} = 10$, $n_{70kDaFN>WT} = 13$ ($P = 0.0032$), $n_{hir} = 6$ ($P = 0.0001$), $n_{YAPS87A>hir} = 10$ ($P = 0.0013$ vs *hir*).

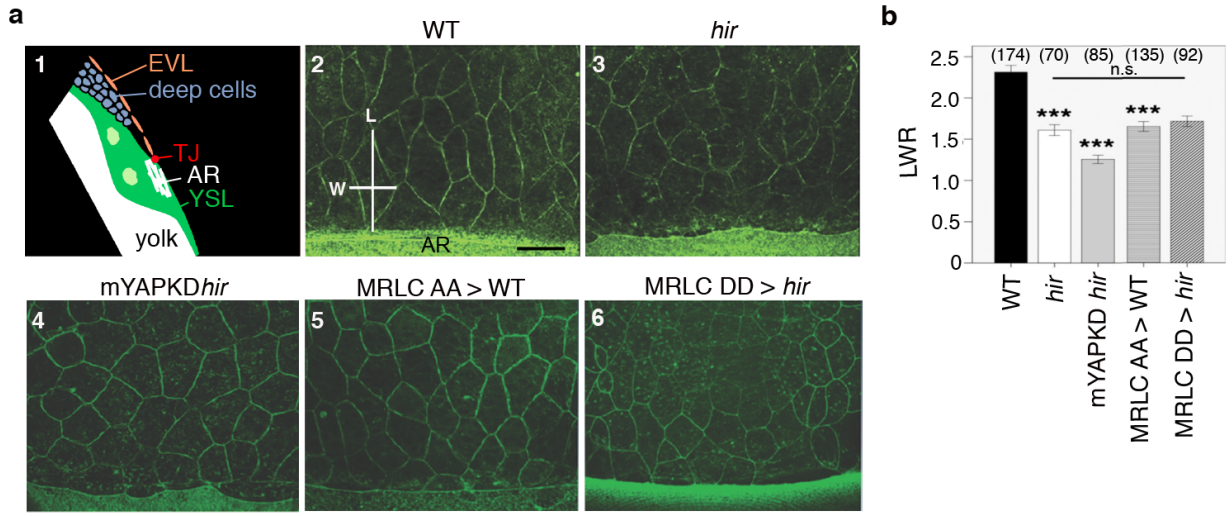
31. Porazinski, S. R., Wang, H. & Furutani-Seiki, M. Essential techniques for introducing medaka to a zebrafish laboratory–towards the combined use of medaka and zebrafish for further genetic dissection of the function of the vertebrate genome. *Methods Mol. Biol.* **770,** 211–241 (2011).
32. Maître, J.-L. *et al.* Adhesion functions in cell sorting by mechanically coupling the cortices of adhering cells. *Science* **338,** 253–256 (2012).
33. Hirose, Y. Varga, Z. M., Kondoh, H. & Furutani-Seiki, M. Single cell lineage and regionalization of cell populations during Medaka neurulation. *Development* **131,** 2553–2563 (2004).
34. Iwanami, N. *et al.* WDR55 is a nucleolar modulator of ribosomal RNA synthesis, cell cycle progression, and teleost organ development. *PLoS Genet.* **4,** e1000171 (2008).
35. Naruse, K. *et al.* A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14,** 820–828 (2004).
36. Itoh, M., Nagafuchi, A., Moroi, S. & Tsukita, S. Involvement of ZO-1 in cadherin-based cell adhesion through its direct binding to alpha catenin and actin filaments. *J. Cell Biol.* **138,** 181–192 (1997).
37. Alexandre, P., Reugels, A. M., Barker, D., Blanc, E. & Clarke, J. D. Neurons derive from the more apical daughter in asymmetric divisions in the zebrafish neural tube. *Nature Neurosci.* **13,** 673–679 (2010).
38. Mayer, M., Depken, M., Bois, J. S., Jülicher, F. & Grill, S. W. Anisotropies in cortical tension reveal the physical basis of polarizing cortical flows. *Nature* **467,** 617–621 (2010).



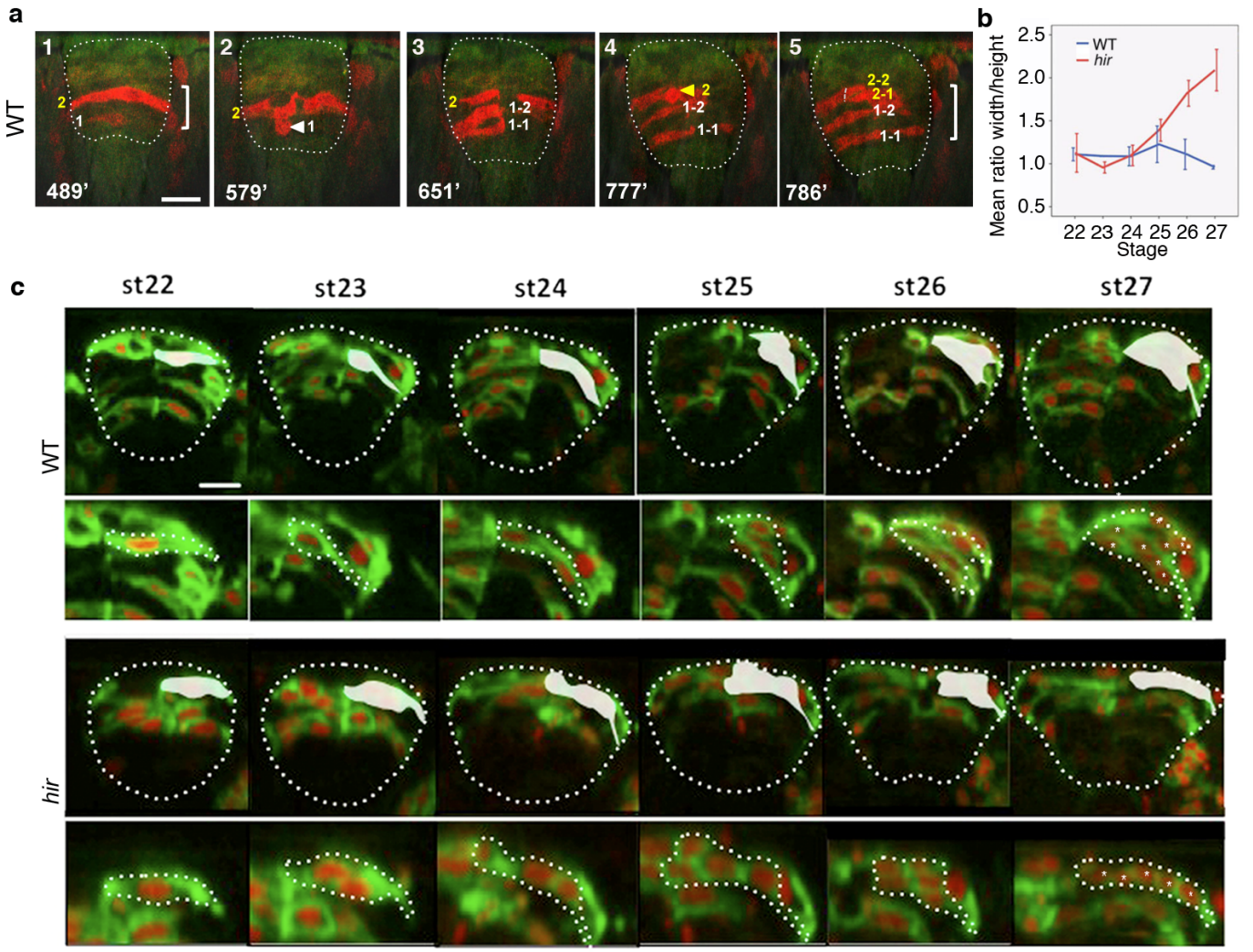
Extended Data Figure 1 | YAP is mutated in *hir* mutants. **a**, *In situ* hybridization of *sox3* showed that the lens placode (arrowhead) is specified in *hir* mutant embryos ($n = 3$) at st. 21. At st. 22, the nascent lens invaginated in WT ($n = 21$), but not in *hir* mutant embryos ($n = 13$, arrowhead). **b**, Two frames from time-lapse imaging of retina of embryos injected with membrane EGFP and nuclear red fluorescent protein (MNFP) mRNAs. In WT ($n = 10$), the nascent lens invaginates from st. 21 (**b1**, margins of the lens indicated by arrowheads with retina to the right), whereas in *hir* ($n = 7$) the lens mostly detached from the retina (**b2′**, arrowheads show lens remnants attached to the retina). Scale bars, 80 μm in **a**; 30 μm in **b**. **c**, Nine recombinants in 1,908 meiosis mapped *hir* close to the YAP gene on chromosome 13 (R: recombinant, C: non-recombinant embryos). **d**, YAP cDNA encodes six protein binding domains/motifs and one transcription activation (TAc) domain; a non-sense mutation in WW1 domain in *hir*. **e**, RT–PCR analysis of YAP mRNA during development. *β-actin* as control. **f**, mRNA of normal YAP and its variants were injected into *hir* mutants. The numbers represent: *hir* phenotype rescue judged via brain thickness, heart migration and Cuvier's duct formation; mutants (judged by genotyping when necessary); survived injected embryos of $hir^{+/-}$ crosses. High dose (400 pg) mRNA of YAP^{hir} variant was injected into WT embryos to examine dominant-negative effects. The rescue by YAP^{4SA} variant required only 20% of the amount required to rescue using normal *YAP* mRNA.



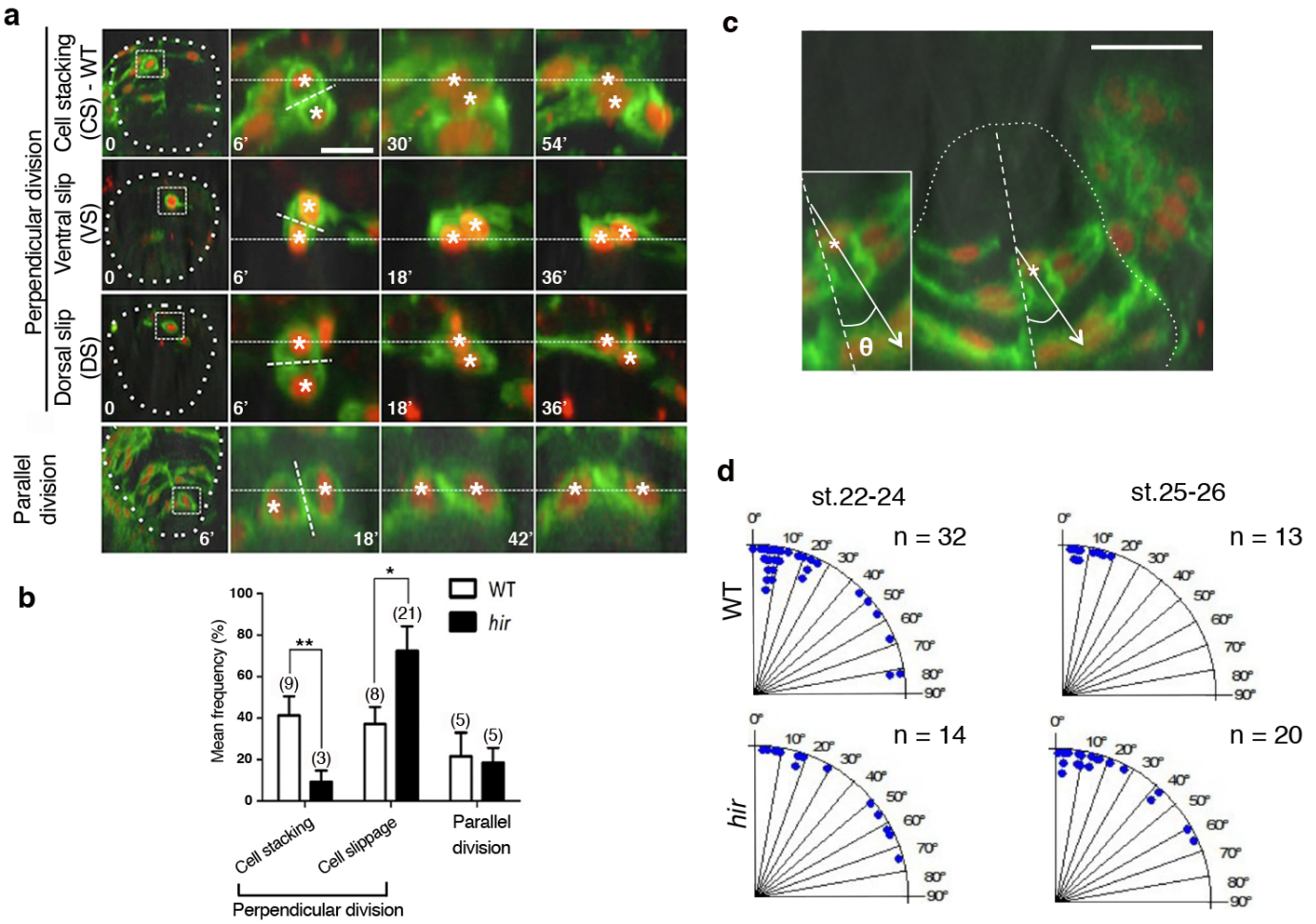
Extended Data Figure 2 | Morpholino knockdown in medaka and zebrafish. **a**, Design of medaka YAP TB and SB MOs relative to translation start (ATG), exons (numbered boxes) and introns. Primers (arrows) used to assess the efficiency of SB MO KD. **b**, Upper panel, proper splicing of YAP transcripts (579 bp) was nearly fully blocked (343 bp, <5% of normal level) by YAP SB MO (5 ng), assessed by RT–PCR; lower panel, *β-actin* control. **c**, WT embryos injected with YAP TB MO and standard control MO. **c1**–**c3** Dorsal and **c1′**–**c3′** lateral views (also Supplementary Table 1). Arrowheads indicate location of heart progenitors. Body flattening and bilateral cardiac progenitor cell migration was affected in a dose-dependent manner. **c2**, **c2′**, Bilateral cardiac progenitor cells fused at the midline but did not migrate anteriorly; **c1**, **c1′** their migration arrests next to the ears at the high dose. The two distinct YAP morpholinos (YAP TB and SB MOs) mimicked the *hir* phenotype (*hir* mutants have a *cardia bifida* phenotype (arrowheads in Fig. 1a1 and a1′)) in a dose-dependent manner. To further verify specificity of the YAP MOs, YAP TB MO was co-injected with human YAP mRNA that does not hybridize with the YAP TB MO. Injection of YAP TB (but not YAP SB) MO into *hir* mutant embryos enhanced the blastopore closure phenotype of *hir* mutants (Fig. 1b, c, Supplementary Table 2). These maternal YAP KD *hir* mutant embryos failed to close the blastopore. Less than half the amount (2 ng) of YAP TB MO was required for causing this phenotype in *hir* mutants compared to that required for WT embryos (5 ng). This blastopore closure phenotype was rescued by medaka YAP mRNA (200 pg) co-injection. **d**–**g**, Zebrafish (ZF) WT embryos injected with three distinct ZFYAP MOs (TB, 5′ UTR and SB) exhibit the blastopore closure phenotype as in medaka (Supplementary Table 3). Efficiencies of ZF YAP and TAZ SB MO KD (1.5 ng each) were assessed by RT–PCR using primers in **d**, **f**, respectively as in **a**, **b**. As reported by Gee *et al.*, co-injection of ZF YAP mRNAs did not rescue the ZF YAP MO phenotype in zebrafish[11]. **h**, Co-injection of ZF TAZ MO (total 2 ng) enhanced slow epiboly of YAP TB KD-injected embryos; control = 89 ± 4.16% ($n = 20$), YAP KD = 70.09 ± 4.7% ($n = 11$), YAP/TAZ KD = 52.5 ± 2.64% ($n = 10$). Error bars show ± s.e.m. ***$P < 0.001$, one-way ANOVA. *y* axis shows percentage epiboly. **i**, **j**, TUNEL for cell death and phosphohistone H3 (PH3) antibody staining for cell proliferation (see methods for sample sizes). Stained cells in the neural tube were counted. Error bars indicate ± s.e.m. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, one-way ANOVA.



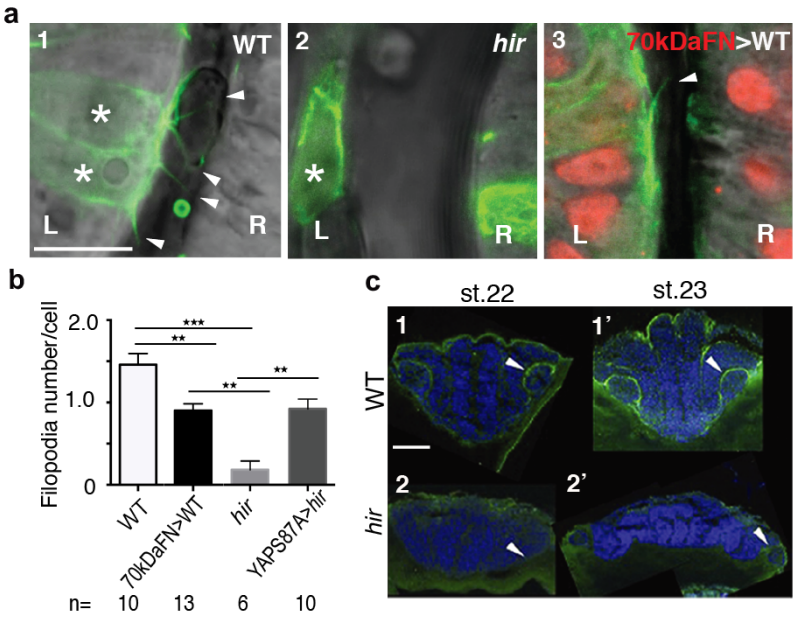
Extended Data Figure 3 | Anisotropic enveloping layer cell shape analysis in *hir* mutants. **a1**, Schematic of sectional view of blastoderm margin of a gastrulating embryo (TJ, tight junction; AR actomyosin ring; YSL, yolk syncytial layer; EVL, enveloping layer); **a2**–**a6**, EVL shape was visualized in phalloidin-stained fixed medaka embryos at 75% epiboly (st. 16, 21 hpf) and compared among, 2 WT ($n = 14$); 3 *hir* ($n = 9$); 4 maternal YAP KD *hir* mutants (mYAPKD*hir*) by TB MO-injection into *hir* embryos ($n = 12$), 5 MRLC-AA (dominant negative form) mRNA-injected WT ($n = 6$); and 6, MRLC-DD (constitutive active form) mRNA-injected *hir* embryos ($n = 4$). **b**, EVL shape anisotropy quantification by the length/width ratio (LWR, shown in **a2**) of marginal EVL cells (up to 4 rows back from the EVL/YSL boundary, shown in Fig. 2d bracket). While EVL shape anisotropy was reduced in *hir* mutant embryos (**a3**) to a level comparable to that of MRLC blocked embryos (**a5**), activation of MRLC in *hir* (**a6**) did not rescue it. Parentheses indicate number of cells measured. Scale bar 30 μm. Error bars represent ± s.e.m. ***$P < 0.001$, one-way ANOVA.



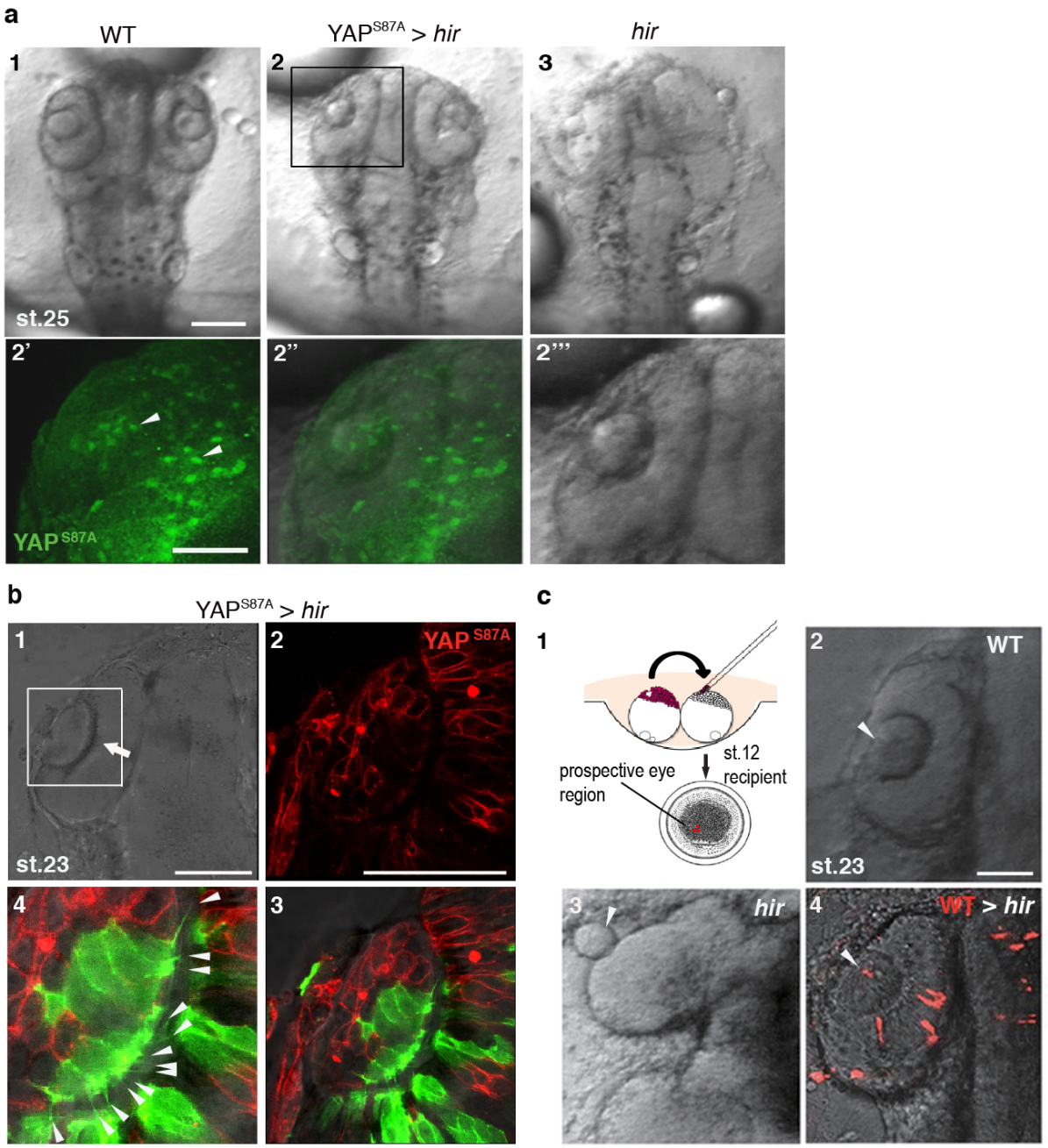
Extended Data Figure 4 | Flattening of the *hir* neural tube is associated with string-like cell arrangements. **a**, Increasing height (indicated by brackets in **a1** and **a5**) of WT neural tube (outlined, $n = 10$) was associated with cell stacking. Time in minutes from st. 21 shown bottom left of each sub-panel. Red fluorescent cells, for example, cell 1 in **a1**, labelled by photo-converting Kaede fluorescent protein, rounded up at the ventricular zone arrowhead in **a2** and divided along the ventricular zone (perpendicular cell division in **a3**) to generate stacked daughter cells 1-1, 1-2, making the neural tube thicker in **a5**. **b**, Width/height ratio of spinal cord, measured from time-lapse imaging of single embryos (WT, *hir* $n = 3$ each), showed that flattening occurred progressively in *hir*. Error bars are ± s.e.m. (see Source Data). **c**, Single-cell tracking of clones (labelled by membrane-GFP and nuclear-RFP) of the growing neural tube at the level of the fifth somite. Lower panels for WT and *hir* show magnified views of shaded regions in upper panels. The flatter and wider neural tube of the *hir* mutant at st. 27 was associated with long chain-like cell arrangements (asterisks, bottom panels of *hir*) tracked from a single neuroepithelial cell at st. 22, compared with the thick cell group generated by cell stacking in WT embryos. Scale bars, 40 μm.



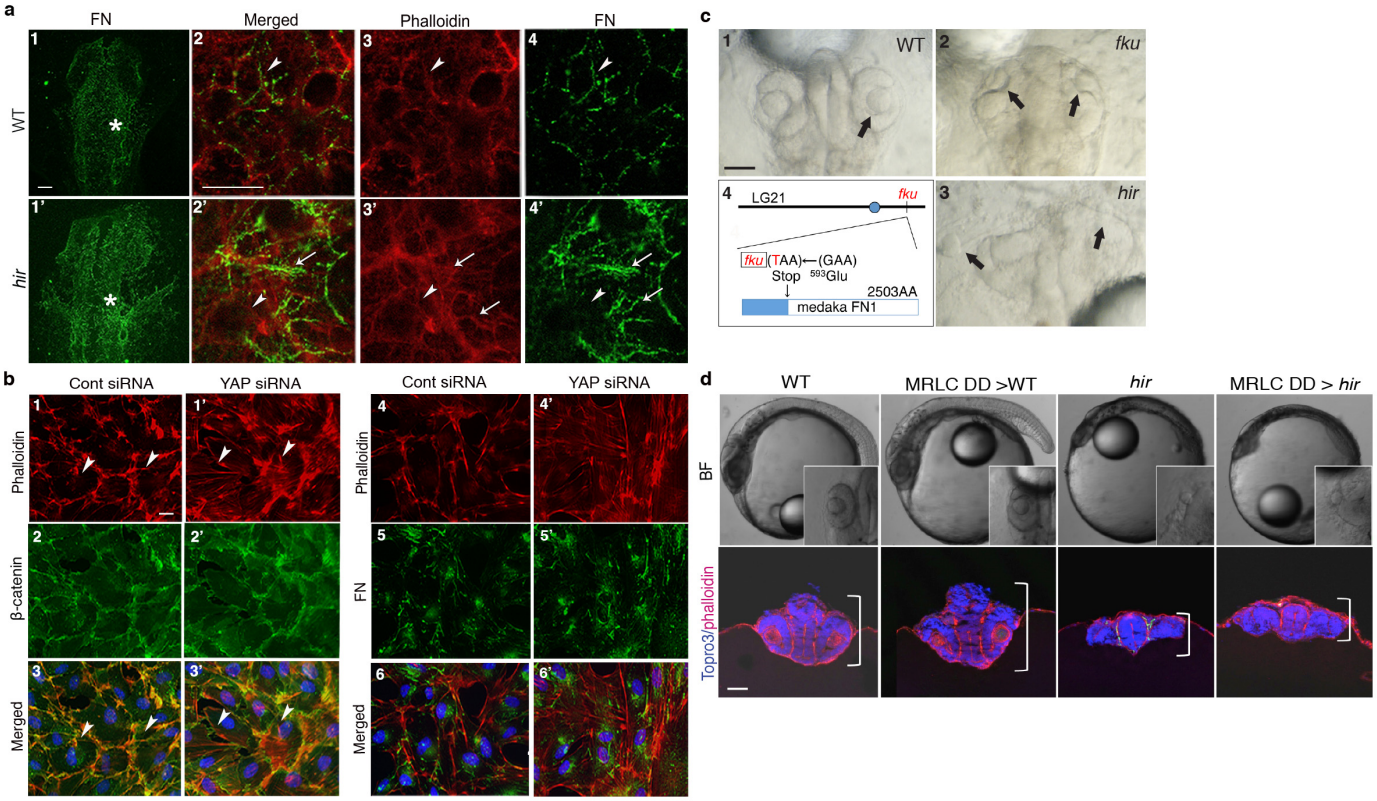
Extended Data Figure 5 | Flattening of the *hir* neural tube is associated with cell stacking failure. **a**–**d**, Single-cell analysis in *hir* neural tube shows cell stacking failure occurred after mitosis (**a**, **b**) and during mitosis (**c**, **d**). Neural progenitor cells divided with spindle orientation 'perpendicular' or 'parallel' to the ventricular zone ('perpendicular' or 'parallel' cell division, respectively). **a**, While daughter cells (asterisks) in WT remained stacked after 45 min following perpendicular cell division (first row), those in *hir* exhibited cell slippage (second and third rows). Telophase neuroepithelial cells in the neural tube, first column; magnified views in second to fourth columns. Dotted lines show division planes. Two types of cell slippage were observed: ventral slippage (VS) where the dorsal daughter cell slipped towards the ventral (second row), and dorsal slippage (DS) where the ventral daughter cell slipped towards the dorsal (third row). After parallel cell division, daughter cells did not change their positions in *hir* (fourth row). **b**, Cell stacking was reduced and cell slippage increased after perpendicular cell division, but cells after parallel cell division remained unaltered in *hir* mutants. Cell numbers in parentheses. Error bars, ± s.e.m. $*P < 0.05$, $**P < 0.01$, *t*-test (see Source Data). **c**, During perpendicular mitosis, daughter cells did not stack properly in *hir* mutants. Cell division orientation (θ) was measured in time-lapse sequences as the acute angle of the telophase cell axis against that of the ventricular zone (for example, dotted line 26° in **c**). **d**, Rose diagrams showing frequency and angle of parallel cell divisions. At st. 25–26 (50–54 hpf) perpendicular cell divisions generated stacked cells against gravitational forces in WT ($n = 3$ embryos at both stages). Far fewer stacked cells were observed in *hir* ($n = 4$ embryos at st. 22–24, $n = 3$ embryos at st. 25–26). These results are illustrated in Fig. 3a. Scale bars, 15 μm in **a**, 40 μm in **c**.



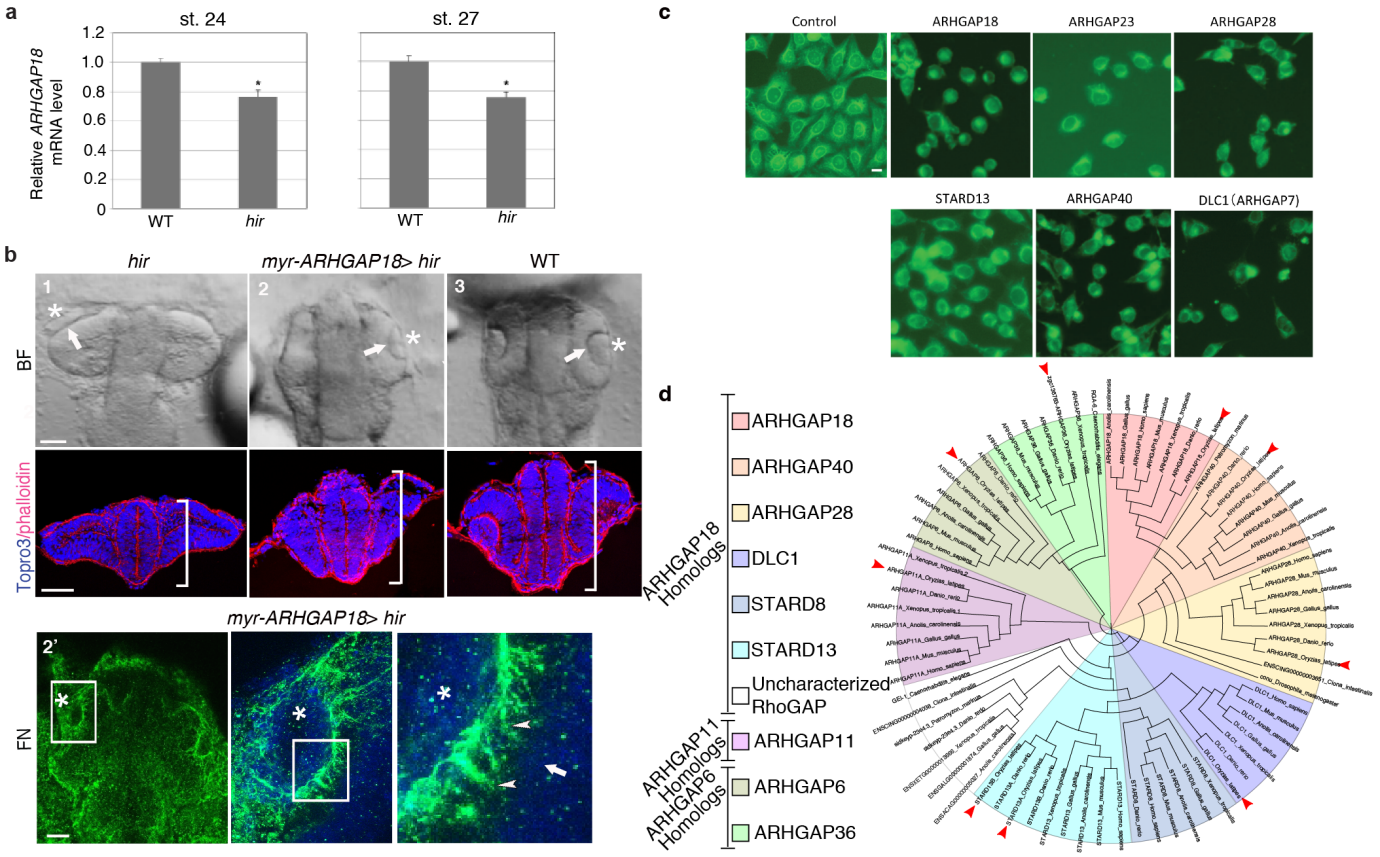
Extended Data Figure 6 | Detachment of lens is associated with loss of filopodia in *hir*. **a**, Representative live images of filopodia (arrowheads) from single lens cells (asterisks) expressing Lifeact–GFP in a mosaic manner; **a1**, WT; **a2**, *hir* and **a3**, 70kDaFN mRNA-injected WT embryos at st. 21.5 when lenses are detaching in *hir* mutants (see Extended Data Fig. 1b for larger views). **a3**, Non-mosaic expression of 70kDaFN mRNA in WT embryos was confirmed by co-injected H2A–red fluorescent protein (RFP) in the nucleus (red). L, lens; R, retina. **b**, Filopodia number per cell was determined (see Extended Data Fig. 7b4 for YAPS87A injected *hir* embryos). *n*, number of analysed embryos. Error bars indicate ± s.e.m. $^{**}P < 0.01$, $^{***}P < 0.001$, one-way ANOVA (Extended Data Fig. 6 Source Data). **c**, Transverse section of integrin-β1 IHC. Strong integrin-β1 localization between lens and retina in st. 22 WT ($n = 2$) (**c1**, arrowhead); no such localization in *hir* ($n = 3$) (**c2**). At st. 23 in *hir* ($n = 3$), weak localization where rounded up lens reattached to retina (**c2′**, arrowhead). Scale bars, 10 μm in **a**; 40 μm in **c**.



Extended Data Figure 7 | The *hir* mutation acts cell non-autonomously. **a**, Mosaic expression of EGFP–YAPS87A by mRNA injection at 16-cell stage in *hir* mutant embryos rescued the *hir* eye phenotype in a2 compared to a1 (WT) and a3 (*hir*). The boxed area in **a2** is magnified in the lower panels (**a2′**–**a2′′′**) fluorescence, merged and bright-field views, respectively. Arrowheads in **a2′** indicate EGFP–YAPS87A-expressing clones. **b**, Non-cell autonomous rescue of filopodia in *hir* mutant lens cells. YAPS87A+ mCherry-CAAX (labels membrane red) mRNA, and Lifeact-EGFP mRNA (labels F-actin green) were injected into different cells at 8–16 cell stage. **b1**, In the invaginated (arrow) *hir* mutant lens (boxed area magnified in **b2** and **b3**, $n = 10$) rescued by mosaic expression of YAPS87A (red), YAPS87A non-expressing mutant cells recovered filopodia (arrowheads in **b4**, magnified view of **b3**). Filopodia number/cell was compared between WT and *hir* in Extended Data Fig. 6b. **c1**, Cells from donor embryos injected with rhodamine (red, top left) were transplanted to a recipient embryo (top right, blastula stage st. 12) at the location fated to be eyes (bottom, animal pole view). **c2**, WT; **c3**, *hir* and **c4**, WT cells transplanted into *hir* mutant eye, causing the lens (arrowhead) to invaginate into the retina as in WT at st. 23 (note that this confocal sectional view represents a fraction of transplanted cells in the whole eye, see Supplementary Table 4 for the frequency of rescue). Scale bars, 40 μm.



Extended Data Figure 8 | F-actin and FN localizations in *hir*. **a**, Whole-mount imaging of WT ($n = 5$) and *hir* ($n = 4$) embryos stained for F-actin (red) and FN (green). **a1**, **a1′**, Whole dorsal view of embryos anterior up, only FN shown; **a2**–**a4**, **a2′**–**a4′**, magnified view of area indicated by asterisks in **a1**, **a1′**; merged **a2**, **a2′**, F-actin **a3**, **a3′** and FN **a4**, **a4′**. Arrowheads indicate cortical F-actin and FN fibrils in WT and corresponding region in *hir* (**a3**, **a4**, **a3′**, **a4′**); arrows show ectopic F-actin aggregates and aberrant FN fibrils in **a3′**, **a4′**. **b**, Immunostaining of 2D cultured RPE1 cells transfected with control (Cont, $n = 21$) and YAP siRNAs ($n = 19$) stained with Phalloidin (**b1**, **b1′**), β-catenin (**b2**, **b2′**) and merged with DAPI (**b3**, **b3′**); phalloidin (**b4**, **4′**), FN (**b5**, **b5′**) and merged with DAPI (**b6**, **b6′**). In marked contrast to the 3D spheroids, FN deposits were not altered in YAP KD cells (**b5**, **b5′**) despite increased F-actin stress fibres (**b1**, **b1′** and **b4**, **b4′**). **c**, The medaka *fku* mutants exhibit lens dislocation (arrows). Live dorsal view of the head of **c1**, WT; **c2**, *fku* and **c3**, *hir* mutant embryos at st. 24. **c4**, The *fku* mutation was mapped to LG21 to the region encompassing the FN1 gene (0 recombinants/1,130 meiosis). Positional cloning identified a non-sense mutation of Glu593 (GAA to TAA) in FN1 (2,503 amino acids). FN1 morpholino KD in WT embryos mimicked the *fku* mutant phenotype. **d**, Constitutive-active MRLC-DD mRNA markedly increased body thickness of WT embryos, but did not rescue the flattened body (brackets in lower panels) and dislocated lens phenotypes of *hir* ($n = 48$). Upper panels, live lateral view (insets, dorsal views of left eyes); lower panels, frontal sections stained with phalloidin (red) and TO-PRO-3 (blue) at st. 25. Scale bars 30 μm, except **a2**, 15 μm and **b1**, 50 μm.



Extended Data Figure 9 | *in vivo* analysis of ARHGAP18 function. **a**, Quantitative RT–PCR analysis showed that *ARHGAP18* mRNA expression in the *hir* mutant is significantly reduced to 76% of WT level. EF1α used as an internal control. Data are shown as means ± s.e.m. ($n = 10$ each; $^*P < 0.001$ Student's *t*-test (two-tailed)). **b**, *myrARHGAP18* mRNA (150 pg) injection rescued the *hir* phenotype (21 rescued/39 *hir*/112 survived embryos). Upper panels, live dorsal view; lower panels, frontal sections stained with phalloidin (red) and TO-PRO-3 (blue) at st. 23; **b1**, uninjected *hir*, **b2**, injected *hir* and **b3**, WT. The lens (asterisk) invaginated into retina (arrows, upper panel) and the neural tube became thicker (brackets in lower panels) in the *myrARHGAP18* mRNA-injected *hir* mutant embryos. **b2′**, FN staining of *myrARHGAP18* mRNA-injected *hir* mutant embryos; boxed area magnified in subsequent panel to the right; invaginated lenses had fine FN fibrils (arrowheads) between lens and retina as in WT (see Fig. 3b1′′). **c**, Phylogenetic analysis identified 16 ARHGAP18 paralogues in vertebrate lineages. Arrowheads show medaka orthologues. **d**, siRNA screening of 40 human ARHGAP genes in HeLa cells showed that KD of five ARHGAP genes exhibited the rounding up phenotype similar to ARHGAP18 inactivation. Scale bars, 30 μm.

LETTER

doi:10.1038/nature14175

Clinical improvement in psoriasis with specific targeting of interleukin-23

Tamara Kopp[1,2]*, Elisabeth Riedl[3]*, Christine Bangert[1]*, Edward P. Bowman[4], Elli Greisenegger[1], Ann Horowitz[4], Harald Kittler[3], Wendy M. Blumenschein[4], Terrill K. McClanahan[4], Thomas Marbury[5], Claus Zachariae[6], Danlin Xu[4], Xiaoli Shirley Hou[4], Anish Mehta[4], Anthe S. Zandvliet[4], Diana Montgomery[4], Frank van Aarle[4] & Sauzanne Khalilieh[4]

Psoriasis is a chronic inflammatory skin disorder that affects approximately 2–3% of the population worldwide and has severe effects on patients' physical and psychological well-being[1–3]. The discovery that psoriasis is an immune-mediated disease has led to more targeted, effective therapies; recent advances have focused on the interleukin (IL)-12/23p40 subunit shared by IL-12 and IL-23. Evidence suggests that specific inhibition of IL-23 would result in improvement in psoriasis. Here we evaluate tildrakizumab, a monoclonal antibody that targets the IL-23p19 subunit, in a three-part, randomized, placebo-controlled, sequential, rising multiple-dose phase I study in patients with moderate-to-severe psoriasis to provide clinical proof that specific targeting of IL-23p19 results in symptomatic improvement of disease severity in human subjects. A 75% reduction in the psoriasis area and severity index (PASI) score (PASI75) was achieved by all subjects in parts 1 and 3 (pooled) in the 3 and 10 mg kg^{-1} groups by day 196. In part 2, 10 out of 15 subjects in the 3 mg kg^{-1} group and 13 out of 14 subjects in the 10 mg kg^{-1} group achieved a PASI75 by day 112. Tildrakizumab demonstrated important clinical improvement in moderate-to-severe psoriasis patients as demonstrated by improvements in PASI scores and histological samples.

Ustekinumab, a monoclonal antibody that binds the IL-12p40 subunit shared by IL-12 and IL-23 (IL-12/23p40), has demonstrated efficacy in the treatment of psoriasis[4,5]. A previous study showed that ustekinumab did not decrease IL-12p35 in human subjects, suggesting that targeting IL-23 was more critical than IL-12 (ref. 6). The IL-23p19 subunit pairs with the IL-12/23p40 subunit to form the heterodimeric cytokine IL-23. Both subunits are overexpressed in psoriasis plaques[7–9]. Human association data and pre-clinical studies strongly suggest that IL-23, not IL-12, drives psoriasis.

We conducted this phase I study in subjects with moderate-to-severe plaque-type psoriasis to evaluate the clinical activity, safety and pharmacokinetics of tildrakizumab (also known as MK-3222 or SCH 900222), a monoclonal antibody that specifically binds the IL-23p19 subunit of IL-23 to neutralize its function, but is unable to bind the IL-12 heterodimer of IL-12/23p40 and IL-12p35 (Extended Data Fig. 1). Evidence shows that IL-23 and its downstream effector cytokines (for example, IL-17A, IL-17F, IL-22 and TNF secreted by αβ T cells (T helper 17 (T_H17), T_H22, T cytotoxic 17 (T_C17), T_C22), γδ T cells, regulatory T cells, natural killer T cells, natural killer cells, type 3 innate lymphoid cells, neutrophils and mast cells) are observed in psoriasis[10–12]. Recent clinical data show that antagonism of downstream cytokine IL-17A has a positive effect in psoriasis, suggesting that an IL-23 antagonist would also be effective[13,14].

Baseline characteristics are shown in Extended Data Table 1. In total, 65 of the 77 enrolled subjects completed the study; 3 were lost to follow up (that is, subjects did not return/ could not be reached; 2 subjects were on 0.1 mg kg^{-1}; 1 subject was on placebo); 2 discontinued owing to adverse events (1 subject on 10 mg kg^{-1} due to convulsions; 1 subject on placebo due to arthralgia and worsening psoriasis); and 7 withdrew consent (1 subject on 0.05 mg kg^{-1}; 2 subjects on 3 mg kg^{-1}; 4 subjects on placebo). In part 1, subjects were randomized to intravenous injections of placebo ($n = 6$) or 0.1 mg kg^{-1} ($n = 3$), 0.5 mg kg^{-1} ($n = 3$), 3 mg kg^{-1} ($n = 6$) or 10 mg kg^{-1} ($n = 6$) tildrakizumab on days 0, 56 and 84. In part 2, subjects were randomized to placebo ($n = 11$) or 3 mg kg^{-1} ($n = 15$) or 10 mg kg^{-1} ($n = 14$) tildrakizumab on days 1, 28 and 56. In part 3, subjects received placebo ($n = 3$) or 0.05 mg kg^{-1} ($n = 6$) or 0.1 mg kg^{-1} ($n = 3$) tildrakizumab on days 1, 56 and 84 (Extended Data Fig. 2).

In this study, disease activity measures were significantly improved after tildrakizumab. Tildrakizumab 0.05–10 mg kg^{-1} resulted in a mean placebo-corrected reduction in PASI score of 50–80% on day 112 with a sustained response at day 196 (Fig. 1). In part 2, a clinically meaningful response (mean placebo-corrected decrease in PASI score of 50%) was still observed on study day 308 (36 weeks after the last administered dose). All 3 and 10 mg kg^{-1} subjects achieved a 75% reduction in the PASI score (PASI75) in part 1 by day 196 and a majority achieved PASI75 in part 2 by day 112 (Extended Data Table 2). A large proportion of 3 and 10 mg kg^{-1} subjects also achieved PASI90 response by day 112 (Extended Data Table 2). A representative subject's clinical photographs demonstrate visual clearance of psoriatic lesions (Fig. 1).

Tildrakizumab was well tolerated up to the maximum dose evaluated (10 mg kg^{-1} intravenously once monthly). The most common adverse effects included headache, nasopharyngitis, upper respiratory infection and cough (Extended Data Table 3). There was no dose-related increase in adverse events or abnormal trends in clinical safety laboratories, vital signs and electrocardiograms (ECGs). There were 11 serious adverse events reported in 8 subjects (Extended Data Table 4). One serious adverse event (convulsions) was deemed to possibly be related to study medication due to the timing of the dose (17 days after dosing with 10 mg kg^{-1}); several confounding factors in this subject include an extended lack of sleep, alcohol consumption and use of benzodiazepine (diazepam), and the subject's attempt to reduce the use of these agents. This phase I trial was limited in scope and sample size; further research will provide a full assessment of safety and tolerability of tildrakizumab. Previous research supports that IL-12 is important for intracellular pathogen immunity and tumour surveillance[15,16]; however, the IL-12/23p40 antagonist ustekinumab has a robust efficacy and safety database[17] and studies with large patient samples will be needed to differentiate selective IL-23 antagonists from broader IL-12/23p40 antagonists in terms of safety.

Tildrakizumab demonstrated slow systemic clearance and a long half-life (~3 weeks). The mean $t_{1/2}$ value ranged from 20.2 to 26.9 days (parts 1, 2 and 3). The mean clearance ranged from 1.57 to 2.50 ml day^{-1} kg^{-1} for 0.05 to 10 mg kg^{-1}. For these pharmacokinetic parameters, no dose trends were observed. For part 2 only, an analysis of variance (ANOVA)

[1]Department of Dermatology, Division of Immunology, Allergy and Infectious Diseases, University of Vienna Medical School, 1090 Vienna, Austria. [2]Juvenis Medical Center, 1010 Vienna, Austria. [3]Department of Dermatology, Division of General Dermatology, University of Vienna Medical School, 1090 Vienna, Austria. [4]Merck & Co., Inc., Whitehouse Station, New Jersey 08889, USA. [5]Orlando Clinical Research Center, Orlando, Florida 32809, USA. [6]Department of Dermato-allergology, Gentofte Hospital, University of Copenhagen, Kildegaardsvej 28, DK-2900 Hellerup, Denmark.
*These authors contributed equally to this work.

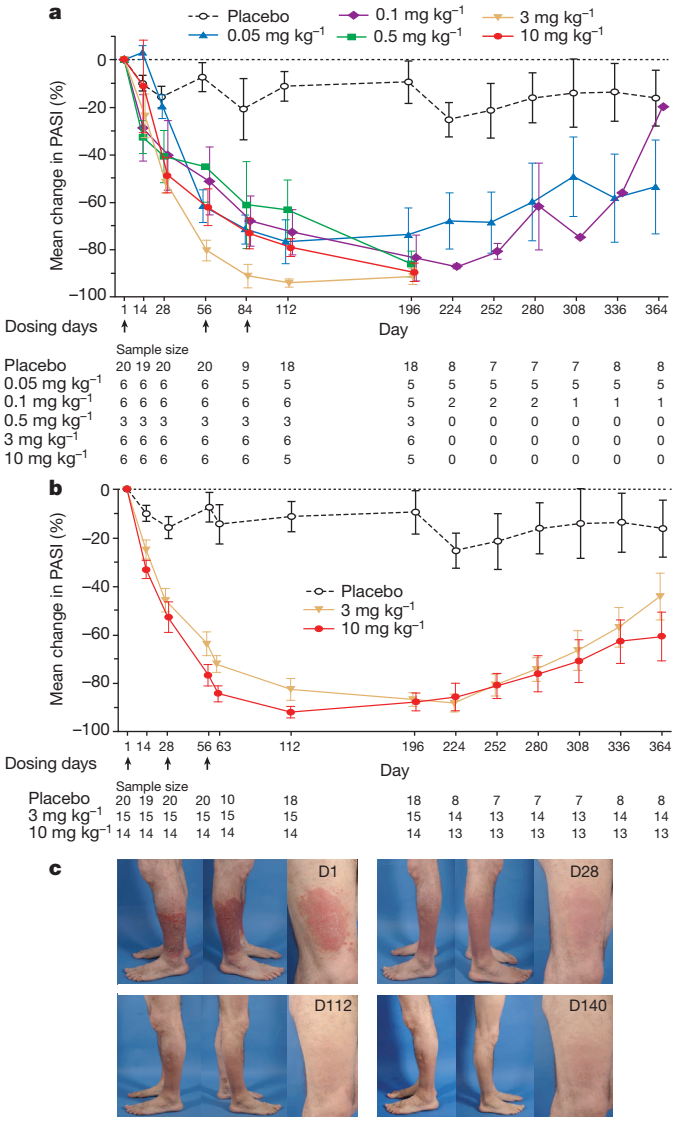


Figure 1 | Tildrakizumab rapidly provides clinical benefit to moderate-to-severe psoriasis patients. **a**, **b**, Mean percentage change in baseline PASI scores in parts 1 and 3 (**a**) and part 2 (**b**) over time. The percentage change in baseline PASI scores in placebo-treated patients from parts 1, 2 and 3 were combined together and used as controls for both graphs. **c**, Representative clinical photographs from a responder in the 10 mg kg^{-1} tildrakizumab group in part 2 ($n = 14$) demonstrate improvement in day 1 (D1) psoriatic lesions at days 28, 112 and140. Error bars denote s.e.m.

dose proportionality test showed that the area under the curve (AUC) and maximum concentration (C_{max}) were dose proportional for doses of 3 and 10 mg kg^{-1}. Exposure was linear over the administered doses.

Of the 56 tildrakizumab-treated subjects, 51 were pre-treatment negative for anti-drug antibodies (ADA). Nine of these (18%) had at least one post-treatment ADA-positive sample and five of these nine showed lower tildrakizumab exposure than ADA-negative subjects. Subjects who had ADA-positive samples did not differ in their PASI response or adverse effect profile; further research will better assess whether the development of ADAs will affect safety or efficacy. Subjects who were ADA-positive before treatment may have had previous treatment with recombinant antibodies that generated a cross-reacting ADA response or there may have been matrix interference in the ADA assay. The screening ADA assay was designed to have a 5% false positive rate to increase assay sensitivity and therefore reduce the risk of missing subjects who develop ADA[18].

We also conducted histological, immunohistochemical and gene expression analyses of lesional skin after tildrakizumab treatment. Psoriatic lesions are characterized by epidermal hyperplasia, hyper-/parakeratosis, hypogranulosis, an increase in suprabasal mitotic keratinocytes, dilation and increased growth of papillary blood vessels, and leukocyte infiltration of both the dermis and epidermis. Both the 3 and 10 mg kg^{-1} doses of tildrakizumab resolved thickened lesional epidermis back to a non-lesional state, but could not entirely convert back to the characteristics of normal skin (Fig. 2a). The amount of CD31$^+$ vessels was increased in lesional skin when compared with non-lesional skin and normal human skin; in contrast to the effects of tildrakizumab on epidermal hyperplasia, only a slight trend to reduced CD31$^+$ vessel density was observed after treatment in lesional skin (Fig. 2b). Blood vascular changes may be underestimated using anti-CD31 staining, due to CD31 expression by lymphatic vessels in the papillary dermis[19].

The histopathological psoriasis severity score (HPSS) was developed based on a previous scoring system and was used to quantify the overall global effect of tildrakizumab on the lesional skin architecture[20]. The total HPSS index was significantly reduced in all tildrakizumab groups, with a mean reduction of 67% (95% confidence interval: 53–81%) (Fig. 2c). In the 3 mg kg^{-1} and 10 mg kg^{-1} tildrakizumab groups, the total HPSS was 16.5 ± 1.3 (mean ± s.e.m.) and 20.0 ± 2.5 before treatment and 6.2 ± 1.9 and 4.4 ± 1.2 after treatment, respectively, and these reductions were statistically significant (Fig. 2c). Although sample sizes for individual groups were small (shown in Fig. 1), significant reductions in epidermal (for example, hyperplasia and mitotic scores), vascular and inflammatory cell infiltrate parameters were observed for the 3 mg kg^{-1} and 10 mg kg^{-1} tildrakizumab subgroups, but not for the placebo group before and after treatment (Fig. 2d). Haematoxylin and eosin photomicrographs are shown in Fig. 3.

IL-23 blockade had a significant effect on the immune compartment of psoriatic plaques. There was consistent clearance of most infiltrating leukocytic cell types in the lesional skin of tildrakizumab patients, suggesting that selective IL-23 blockade may induce disease remission. Consistent with a normalized skin environment, the proliferation marker Ki67 (Fig. 2e, f) and the abnormally expressed epithelial antigen keratin 16 (Fig. 2g, h) were normalized after treatment. This correlates with a significant reduction in mitosis within the suprabasal epidermal layer (Fig. 2d). The inflammatory infiltrate in psoriatic lesions consists of CD3$^+$, CD4$^+$ and CD8$^+$ T cells; CD68$^+$ myeloid cells; BDCA-2$^+$ plasmacytoid dendritic cells; CD11c$^+$ myeloid dendritic cells; and CD15$^+$ neutrophils. Epidermal CD4$^+$ and CD8$^+$ T cells (Fig. 4a, b), dermal myeloid dendritic cells (Fig. 4c), plasmacytoid dendritic cells (Fig. 4d), and CD15$^+$ neutrophils (Fig. 4f) were significantly decreased following tildrakizumab dosing. Importantly, dermal CD4$^+$ and CD8$^+$ T cells (data not shown) demonstrated decreased levels. Langerin$^+$ epidermal Langerhans cell density was minimally altered in lesional psoriatic skin and the density was not statistically affected by tildrakizumab treatment (Fig. 4e).

Because IL-23 is a master cytokine in the orchestration of cutaneous inflammation[13], we performed detailed analysis of lesional skin in 10 mg kg^{-1} tildrakizumab and placebo patients to identify the cellular source of IL-23p19 in the dermal compartment. Results demonstrate CD11c$^+$ myeloid dendritic cells, CD15$^+$ neutrophils, and CD163$^+$ macrophages to be the main producers of IL-23p19, with a negligible number of CD117$^+$ mast cells (Fig. 5). These findings are consistent with previous reports showing that IL-23 is expressed by activated human macrophages and dendritic cells[8]; these data complement messenger RNA data showing overexpression of IL-23p19 and IL-12/23p40 in psoriatic skin[9,10].

Importantly, IL-23p19 expression trended to almost completely disappear after tildrakizumab treatment (Fig. 5), although it was not significantly different owing to high variability in infiltrating cell counts and the small sample size. The reduction in IL-23p19$^+$ cells after tildrakizumab treatment has three potential explanations. First, tildrakizumab still

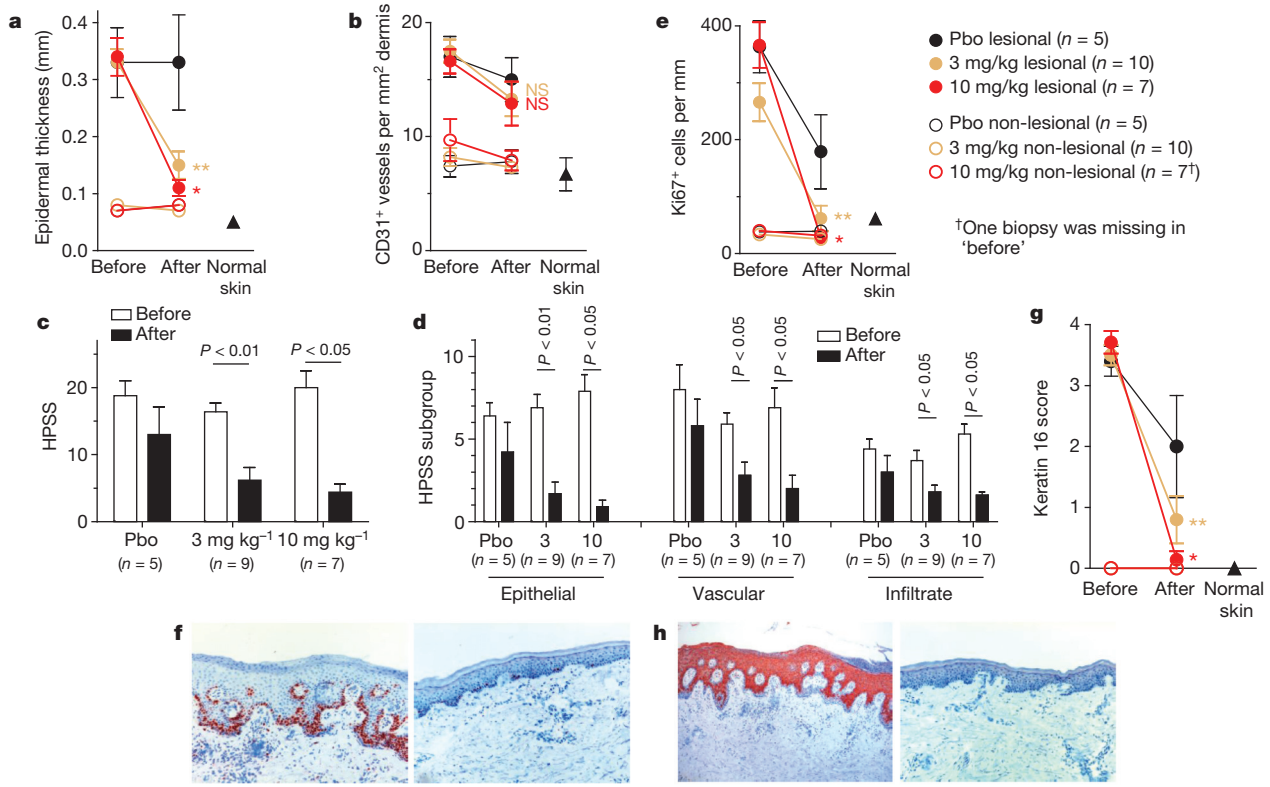


Figure 2 | Tildrakizumab resolves altered epithelium in psoriasis patients. **a**, **b**, Lesional (filled) and non-lesional (open) biopsies were taken before and after three doses. Skin biopsies were taken from healthy volunteers (normal skin). Epithelial thickness (**a**) or dermal $CD31^+$ vessel number (**b**) was measured in sectioned biopsies. **c**, **d**, HPSS scoring graded for overall epidermis, vascular and infiltrate changes (**c**) and individual scores for each parameter (**d**). **e**–**h**, Representative photomicrographs and analysis from lesional skin of a responder in part 2 in the 10 mg kg^{-1} tildrakizumab group ($n = 7$ for subjects on 10 mg kg^{-1} who provided skin samples) before (left) and after (right) dose presented for anti-Ki67 (**e**, **f**) and anti-keratin 16 (**g**, **h**). Biopsies analysed for placebo (pbo) ($n = 5$), 3 mg kg^{-1} ($n = 10$) and 10 mg kg^{-1} tildrakizumab ($n = 7$), and healthy volunteers ($n = 5$). *$P < 0.05$, *$P < 0.01$ (Wilcoxon matched-pairs signed-rank test). Data are mean ± s.e.m. Original magnifications, ×100.

present in the biopsy specimen binds and blocks the IL-23-binding site of the anti-IL-23p19 monoclonal antibody used for immunohistochemical detection. Second, tildrakizumab exhibits a direct effect on the IL-23-producing cells resulting in their deletion. Third, the decrease in IL-23-producing cells is not an immediate direct downstream effect of IL-23 blockade, but rather a general sign of the healing plaque. Importantly, cross-blocking experiments demonstrated that tildrakizumab does not interfere with the ability of the detecting anti-IL-23p19 antibody to recognize and bind IL-23p19 (L. Bald, personal communication), supporting one of the latter two explanations.

Gene expression analysis of lesional skin pre- and post-treatment was performed to correlate histological and immunophenotypical changes with tildrakizumab effects on skin gene expression. IL-19 and IL-20 are keratinocyte growth/differentiation factors that were indirectly increased in skin by IL-23 action in preclinical models[10,21,22]. Tildrakizumab decreased IL-19 and IL-20 message levels that correlated with decreased epithelial thickness, decreased numbers of $Ki67^+$ proliferating keratinocytes, and normalized keratinocyte differentiation (that is, keratin 16). The anti-microbial factors S100A7 and LCN2 were indirectly increased in skin by IL-23 action in preclinical models[10,23,24] and tildrakizumab also normalized their expression. T_H17 CD4 T cells express the chemokine receptor CCR6, whose ligand CCL20 is overexpressed in lesional skin. Neutrophils express CXCR1 and CXCR2, whose ligand CXCL8/IL-8 is also overexpressed[25,26]. Tildrakizumab decreased CCL20 and CXCL8/IL-8 message levels, which correlated with reduced CD4 T-cell and neutrophil infiltrates after tildrakizumab dosing (Fig. 6).

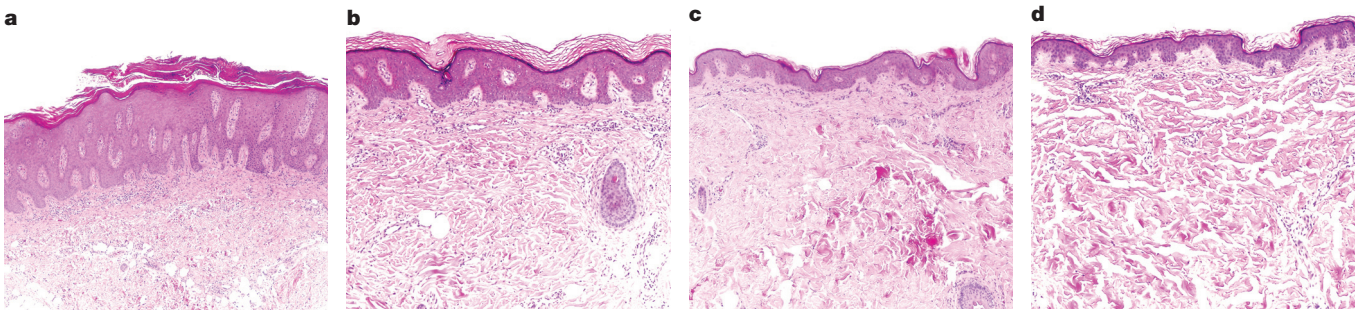


Figure 3 | Representative haematoxylin and eosin photomicrographs of sections from lesional and non-lesional skin. **a**, **b**, Lesional skin before (**a**) and after (**b**) treatment. **c**, **d**, Non-lesional skin before (**c**) and after (**d**) treatment. Sections represent a subject who was a responder in part 2 in the tildrakizumab dose group ($n = 7$). Original magnification, ×200.

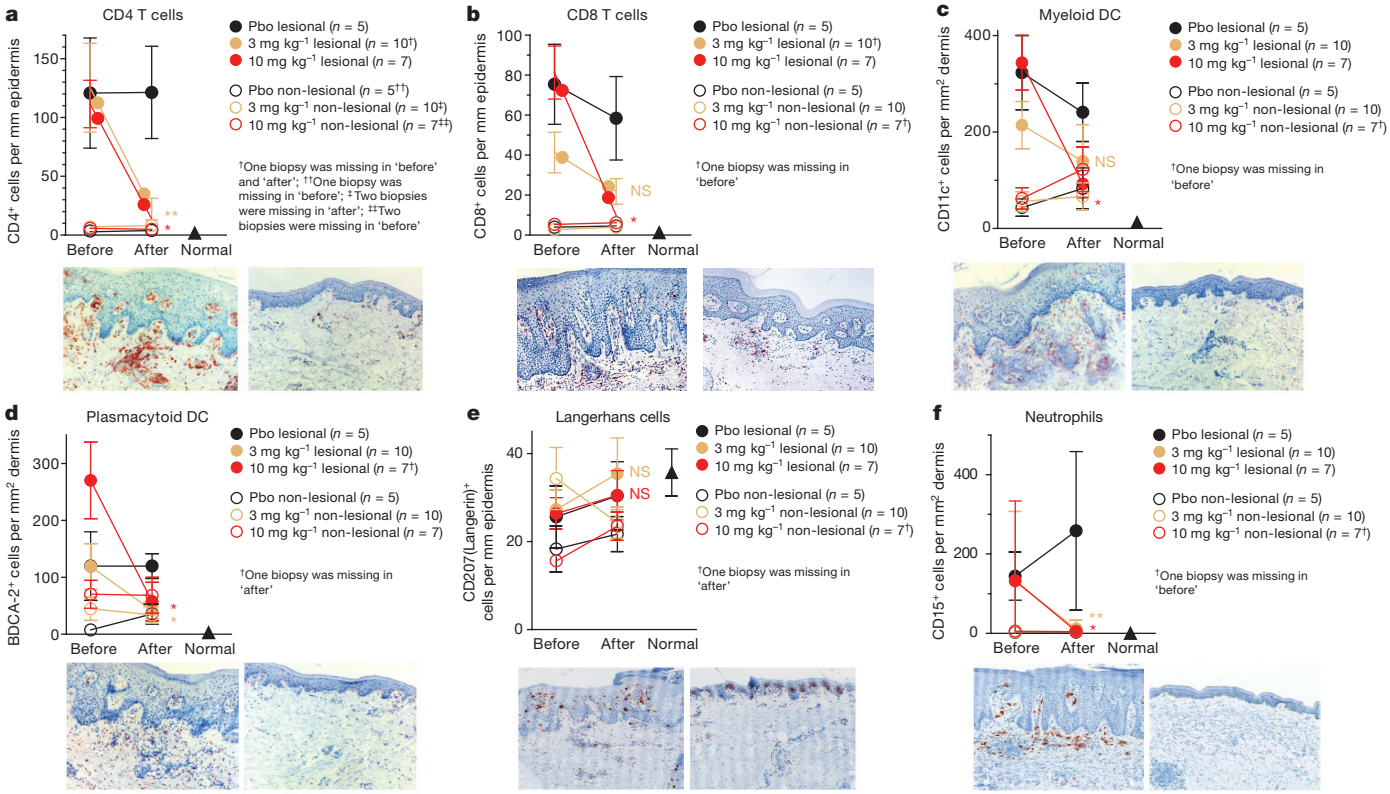


Figure 4 | Tildrakizumab diminishes the elevated dermal infiltrate in psoriatic skin. **a**–**f**, Lesional (filled) and non-lesional (open) biopsies were taken before and after three tildrakizumab doses or placebo. Skin biopsies were taken from healthy volunteers (normal skin). Biopsies were stained with anti-CD4 (**a**), anti-CD8 (**b**), anti-CD11c (**c**), anti-BDCA-2 (**d**), anti-CD207/Langerin (**e**) or anti-CD15 (**f**). Representative photomicrographs of lesional skin from a responder in the 10 mg kg^{-1} group in part 2 ($n = 7$ for subjects on 10 mg kg^{-1} who provided skin samples) showing pre-dose (left) and post-dose (right) are presented. Biopsies were analysed from patients on placebo ($n = 5$), 3 mg kg^{-1} ($n = 10$) and 10 mg kg^{-1} ($n = 7$) tildrakizumab, and healthy volunteers ($n = 5$). DC, dendritic cells. $*P < 0.05$, $**P < 0.01$ (Wilcoxon matched pairs signed-rank test); NS, not significant. Data are mean ± s.e.m. Original magnifications, ×100.

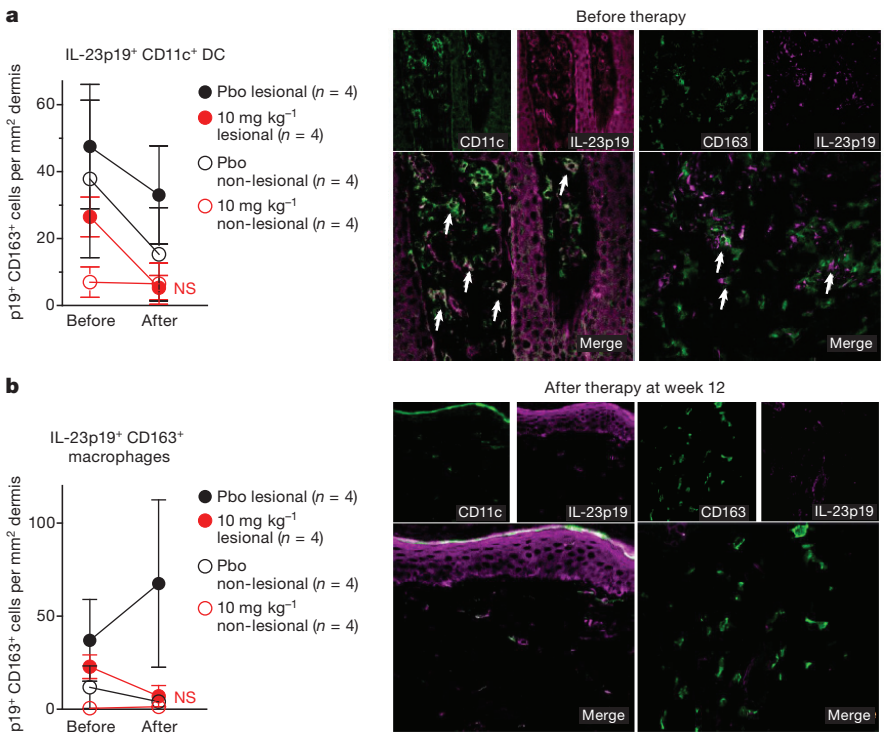


Figure 5 | Identification of IL-23p19-producing cells in psoriatic lesions. **a**, **b**, Four-millimetre lesional (filled circles) and non-lesional (open circles) punch biopsies were taken before dosing and after a course of three tildrakizumab doses or placebo. Skin biopsy sections were co-stained with anti-IL-23p19 and anti-CD11c (**a**) or anti-IL-23p19 and anti-CD163 (**b**). TissueQuest software was applied for image cytometry of immunofluorescence stainings. Skin biopsies were analysed for placebo ($n = 4$) and 10 mg kg^{-1} tildrakizumab ($n = 4$). Representative photomicrographs of lesional skin from a responder in the 10 mg kg^{-1} group in part 2 ($n = 7$ for subjects on 10 mg kg^{-1} who provided skin samples). Data are mean ± s.e.m. Original magnifications, ×400.

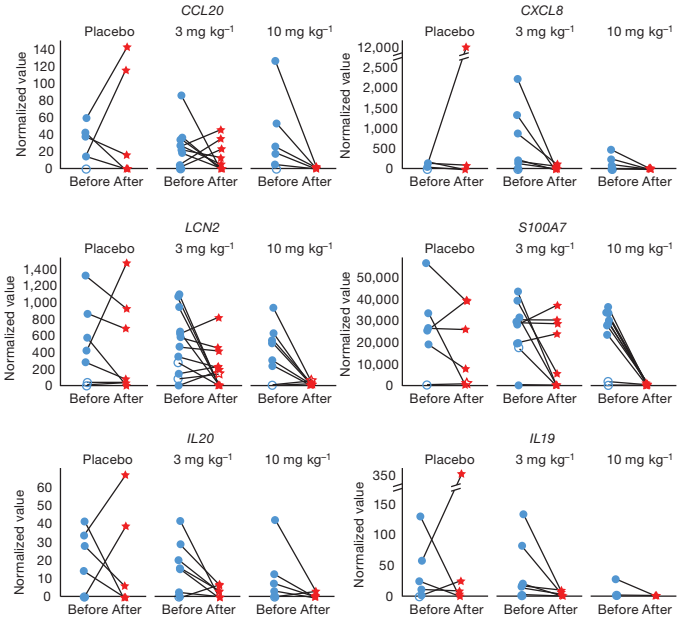


Figure 6 | Tildrakizumab inhibits IL-23-associated psoriatic skin gene expression. Four-millimetre lesional (filled circles) and non-lesional (open circles) punch biopsies were taken before dosing and after three tildrakizumab doses or placebo. Lysates were prepared from biopsy sections for analysis on the NanoString nCounter gene expression platform. Skin biopsies were analysed from patients on placebo ($n = 5$), 3 mg kg^{-1} tildrakizumab ($n = 10$) or 10 mg kg^{-1} tildrakizumab ($n = 7$). All genes were detectably increased in lesional skin compared with non-lesional skin, and were significantly decreased by at least one tildrakizumab dose using a $P < 0.05$ criteria in the Wilcoxon matched-pairs signed-rank test.

Subsequent studies on tildrakizumab will address limitations in the study design. Although our data indicate dose-related efficacy, this trial did not have sufficient sample size to clearly show a dose response. A randomized, double-blind, placebo-controlled, parallel-group dose-range finding phase IIb trial will identify optimal doses (ClinicalTrials.gov (https://clinicaltrials.gov/) identifier NCT01225731). Skin biopsies were collected and processed for histological and immunohistochemistry analyses, but biopsies were not collected specifically for gene expression analysis. Dedicated biopsy collection for gene expression analysis will be performed in NCT01225731 to evaluate modulation of additional IL-23 pathway genes (for example, IL-17A, IL-17C, IL-17F and IL-22) that were below the limit of detection for technology used in this study to recover message from formalin-fixed tissue. Lastly, this trial was not designed to perform comparisons with other biologic treatments; an active comparator trial with a currently available treatment is currently in progress (ClinicalTrials.gov identifier NCT01729754).

In summary, this study demonstrates that IL-23 is crucial in psoriasis pathogenesis and that selectively inhibiting IL-23 is a potentially important treatment modality. Tildrakizumab provided important clinical improvement and was found to be well tolerated in these patients with moderate to severe psoriasis. Further development of tildrakizumab is warranted based on these results to determine whether selective targeting of IL-23 can provide similar or better efficacy while reducing safety concerns that are associated with other biologic agents currently approved for the treatment of psoriasis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 December 2013; accepted 23 December 2014.
Published online 9 March; corrected online 13 May 2015 (see full-text HTML version for details).

1. Schön, M. P. & Boehncke, W. H. Psoriasis. *N. Engl. J. Med.* **352,** 1899–1912 (2005).
2. Lowes, M. A., Bowcock, A. M. & Krueger, J. G. Pathogenesis and therapy of psoriasis. *Nature* **445,** 866–873 (2007).
3. Krueger, G. *et al.* The impact of psoriasis on quality of life: results of a 1998 National Psoriasis Foundation patient-membership survey. *Arch. Dermatol.* **137,** 280–284 (2001).
4. Leonardi, C. L. *et al.* Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 76-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 1). *Lancet* **371,** 1665–1674 (2008).
5. Papp, K. A. *et al.* Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2). *Lancet* **371,** 1675–1684 (2008).
6. Toichi, E. *et al.* An anti-IL-12p40 antibody down-regulates type 1 cytokines, chemokines, and IL-12/IL-23 in psoriasis. *J. Immunol.* **177,** 4917–4926 (2006).
7. Oppmann, B. *et al.* Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* **13,** 715–725 (2000).
8. Lee, E. *et al.* Increased expression of interleukin 23 p19 and p40 in lesional skin of patients with psoriasis vulgaris. *J. Exp. Med.* **199,** 125–130 (2004).
9. Chan, J. R. *et al.* IL-23 stimulates epidermal hyperplasia via TNF and IL-20R2-dependent mechanisms with implications for psoriasis pathogenesis. *J. Exp. Med.* **203,** 2577–2587 (2006).
10. Lowes, M. A., Suárez-Fariñas, M. & Krueger, J. G. Immunology of psoriasis. *Annu. Rev. Immunol.* **32,** 227–255 (2014).
11. Villanova, F. *et al.* Characterization of innate lymphoid cells in human skin and blood demonstrates increase of NKp44^{+} ILC3 in psoriasis. *J. Invest. Dermatol.* **134,** 984–991 (2014).
12. Keijsers, R. R. *et al.* Balance of Treg vs. T-helper cells in the transition from symptomless to lesional psoriatic skin. *Br. J. Dermatol.* **168,** 1294–1302 (2013).
13. Krueger, J. G. *et al.* IL-17A is essential for cell activation and inflammatory gene circuits in subjects with psoriasis. *J. Allergy Clin. Immunol.* **130,** 145–154 (2012).
14. Hueber, W. *et al.* Effects of AIN457, a fully human antibody to interleukin-17A, on psoriasis, rheumatoid arthritis, and uveitis. *Sci. Transl. Med.* **2,** 52ra72 (2010).
15. Ngiow, S. F., Teng, M. W. & Smyth, M. J. A balance of interleukin-12 and -23 in cancer. *Trends Immunol.* **34,** 548–555 (2013).
16. Bustamante, J., Picard, C., Boisson-Dupuis, S., Abel, L. & Casanova, J. L. Genetic lessons learned from X-linked Mendelian susceptibility to mycobacterial diseases. *Ann. N. Y. Acad. Sci.* **1246,** 92–101 (2011).
17. Papp, K. A. *et al.* Long-term safety of ustekinumab in patients with moderate-to-severe psoriasis: final results from 5 years of follow-up. *Br. J. Dermatol.* **168,** 844–854 (2013).
18. United States Food and Drug Administration. Guidance for Industry; Assay Development for Immunogenicity Testing of Therapeutic Proteins. Available at http://www.fda.gov/downloads/Drugs/Guidances/UCM192750.pdf (2009).
19. Torzicky, M. *et al.* Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31) and CD99 are critical in lymphatic transmigration of human dendritic cells. *J. Invest. Dermatol.* **132,** 1149–1157 (2012).
20. Trozak, D. J. Histologic grading system for psoriasis vulgaris. *Int. J. Dermatol.* **33,** 380–381 (1994).
21. Wolk, K. *et al.* The Th17 cytokine IL-22 induces IL-20 production in keratinocytes: a novel immunological cascade with potential relevance in psoriasis. *Eur. J. Immunol.* **39,** 3570–3581 (2009).
22. Sa, S. M. *et al.* The effects of IL-20 subfamily cytokines on reconstituted human epidermis suggest potential roles in cutaneous innate defense and pathogenic adaptive immunity in psoriasis. *J. Immunol.* **178,** 2229–2240 (2007).
23. Guttman-Yassky, E. *et al.* Low expression of the IL-23/Th17 pathway in atopic dermatitis compared to psoriasis. *J. Immunol.* **181,** 7420–7427 (2008).
24. Liang, S. C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203,** 2271–2279 (2006).
25. Homey, B. *et al.* Up-regulation of macrophage inflammatory protein-3 α/CCL20 and CC chemokine receptor 6 in psoriasis. *J. Immunol.* **164,** 6621–6632 (2000).
26. Lew, W., Lee, E. & Krueger, J. G. Psoriasis genomics: analysis of proinflammatory (type 1) gene expression in large plaque (Western) and small plaque (Asian) psoriasis vulgaris. *Br. J. Dermatol.* **150,** 668–676 (2004).

Acknowledgements The authors thank E. Marcantonio (Merck & Co., Inc.) for his supervision of research and review of this manuscript and J. Pawlowski (Merck & Co., Inc) for editorial and administrative assistance with this manuscript. This study was funded by Merck & Co., Inc.

Author Contributions All authors provided substantive suggestions and critically reviewed the manuscript. T.K., E.R., C.B., A.M., E.P.B. and S.K. wrote sections of the initial draft. T.K., E.R., C.B., E.P.B., H.K., T.K.M., A.S.Z., W.M.B., X.S.H., D.M. and S.K. collected/assembled data and/or performed or supervised analyses. T.K., A.H., D.X. and X.S.H. were involved in the design and planning of the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.K. (sauzanne.khalilieh@merck.com).

METHODS

This study (protocol 5382) was conducted from October 2008 to September 2011 in nine study sites. The protocol was reviewed and approved by local Institutional Review Boards; all subjects signed written informed consent. Subjects also provided consent to publication of clinical photographs. The study followed principles of Good Clinical Practice. Skin biopsies from healthy volunteers were obtained at the Vienna University Medical School. The investigational protocol was approved by the local Ethics committee of Vienna University's Medical School, the Midlands Institutional Review Board (USA), and Ethical Committee for the Region of Capital (Denmark). All subjects agreed to study terms by giving their informed consent including their approval of documentation and publication of clinical pictures.

Subjects. Subjects who were included in this study were aged 18–65 years and were to have a diagnosis of established and stable psoriasis vulgaris with an off-treatment, screening PASI score no more than 40% different than baseline PASI. Subjects were to have moderate to severe psoriasis that was severe enough to render them eligible for systemic therapy (that is, body surface area was $\geq$10% and PASI was $\geq$12 (if on therapy body surface area was $\geq$10% and PASI was $\geq$10)). Target lesions were on the head, trunk, arms or legs and were $\geq$10 cm^2 ($\sim$3 cm diameter). The total numerical ratings for erythema, infiltration and desquamation were $\geq$6 out of possible 12 with 0 = none, 1 = mild, 2 = moderate, 3 = severe, 4 = very severe; desquamation must have been $\geq$2.

All subjects were to be otherwise healthy: normal ECG, normal vital signs, no history of latent or active tuberculosis; subjects were to be negative for tuberculosis 1 month before and had no recent contact with tuberculosis, no history of malignancy, no inflammatory bowel disease, ulcers, gastrointestinal bleeding, gastrointestinal tract surgery, pancreatic injury, liver disease, impaired renal function, urinary obstruction, HIV, hepatitis C, hepatitis B, no infectious disease $<$4 weeks before study drug administration, no blood donation $<$60 days before study, no vaccine $<$1 month before the study, no immunosuppressive agent $<$4 weeks before study, no phototherapy $<$4 weeks before study, no topical treatments $<$1 week before study, and no monoclonal antibodies $<$3 months before study.

Study design. This was a three-part, multiple-dose, randomized, placebo-controlled, patient- and evaluator-blind, multi-centre study of tildrakizumab in subjects with psoriasis (Extended Data Fig. 2). Part 1 was a safety, tolerability and pharmacokinetic study that examined rising doses of tildrakizumab and was planned to enroll 24 subjects with psoriasis into four ascending dose cohorts. Subjects were randomized within each cohort in a 3:1 ratio of active to placebo. Each subject received a single dose of 0.1 mg kg^{-1} (cohort 1), 0.5 mg kg^{-1} (cohort 2), 3 mg kg^{-1} (cohort 3) or 10 mg kg^{-1} (cohort 4) of tildrakizumab intravenously administered over 1 h or placebo. After an 8-week safety observation period, each subject in part 1 (cohorts 1 to 4) received an additional two doses of tildrakizumab intravenously at 0.1, 0.5, 3 or 10 mg kg^{-1} or placebo, respectively, 4 weeks apart. Serum pharmacokinetic samples were obtained pre-dose and at the time points specified in Evaluation Criteria. Subjects were followed for 196 days after the first dose.

Part 2 examined the efficacy of 3 and 10 mg kg^{-1} tildrakizumab with a higher number of subjects per group and was planned to enroll 40 subjects with psoriasis into two ascending dose cohorts (cohorts 5 and 6). Subjects were randomized in a 3:1 ratio of active to placebo. Subjects in cohort 5 received three doses of 3 mg kg^{-1} of tildrakizumab intravenously or placebo given at weeks 0, 4 and 8. Cohort 6 received three doses of 10 mg kg^{-1} tildrakizumab intravenously or placebo given at weeks 0, 4 and 8 (Extended Data Fig. 2). Serum pharmacokinetic samples were obtained pre-dose and at the intervals specified in Evaluation Criteria. Subjects were followed for 364 days after the first dose. Skin biopsies of lesional and non-lesional skin for histology and immunohistochemistry evaluation were obtained at baseline (before) and after a course of three doses of tildrakizumab or placebo at either day 63 or day 84 in a subgroup of patients at two clinical sites.

Part 3 was added owing to efficacy observed at the low doses in part 1; we further examined efficacy using data from part 3. Part 3 of the study was planned to enroll 12 subjects with psoriasis into cohort 7; these patients were randomized in a ratio of 2:1:1 to receive 0.05 or 0.1 mg kg^{-1} tildrakizumab or placebo. All three treatments were administered in a parallel fashion (that is, there was no sequential dose escalation). Subjects received three doses of 0.05 or 0.1 mg kg^{-1} of tildrakizumab or placebo intravenously given at weeks 0, 8 and 12 (Extended Data Fig. 2). Serum pharmacokinetic samples were obtained pre-dose and at the intervals specified in Evaluation Criteria. Subjects were followed for 364 days after the first dose. Skin biopsies of lesional and non-lesional skin for histology and immunohistochemistry evaluation were obtained at baseline and on day 84.

PASI assessments were carried out at protocol-specified intervals until day 196 in part 1 and day 364 in parts 2 and 3.

Before administration of study medication, a computer-generated randomization number was assigned to all subjects who met the study's entry criteria. The randomization number was sent to the study site by the sponsor and was then used to assign treatments to subjects. Blinding was conducted under third-party blinding conditions in which only the site pharmacist, who was not involved with study procedures or evaluations, had knowledge of the treatment code. The pharmacist prepared the treatment, which was then dispensed in a blinded fashion by the investigator. If blinding needed to be broken, the reason was to be recorded and the sponsor notified immediately. Blinded assessments were done by the sponsor and investigators on safety, pharmacodynamics, and pharmacokinetics. The randomization code was sent to a bioanalytical unit for analysis of samples.

Objectives. The primary objectives were to determine the safety, tolerability and pharmacokinetics of rising intravenous doses of tildrakizumab in subjects with moderate-to-severe psoriasis (PASI $\geq$ 12). The secondary objective was to evaluate the effect of tildrakizumab on skin disease in moderate-to-severe psoriasis; although change in PASI was a secondary end point, the study was powered to detect a placebo-corrected change in PASI of 30%. The primary end point was the placebo-corrected percentage change in PASI at day 112 (week 16). Additional endpoints included skin biopsies for exploratory immunohistochemistry, histology and gene expression.

Timing of dose administration and evaluations. In parts 1 and 3, treatments were administered on study days 1, 56 and 84. Biopsies were collected at baseline (within 3–7 days of dosing on day 1) and on study day 112 (4 weeks after the last dose). PASI scores were determined at screening (8–56 days before dose), baseline (3–7 days before the dose), study days 14, 28, 56, 84 and 112 (4 weeks post the last dose), 196 and 364 (part 3 only).

In part 2, study drug was administered on study days 1, 28 and 56. Skin biopsies were collected at baseline (3–7 days before drug administration on day 1) and study day 84 (4 weeks post the last dose). PASI scores were evaluated at screening (day 8 to 56 days before study drug administration on day 1), baseline (as defined for part 1/3) and study days 14, 28, 56, 63 and 112 (8 weeks after the last dose), 196, 224, 252, 280, 308, 336 and 364.

Beyond study day 7, a 3-day window was allowed for scheduled visits. This window was considered acceptable owing to the exploratory, proof-of-concept nature of the study, the 21-day half-life of tildrakizumab, and the gradual time course for skin clearing.

Safety and tolerability. Adverse events were recorded and vital signs were monitored throughout the study. Adverse events were evaluated for severity and relationship to study drug by the investigators. The decision to discontinue therapy due to adverse events was made by the investigator. Safety assessments (including ECGs, clinical laboratory measurements, adverse event monitoring, collection of vital signs), and blood and urine assessments were performed throughout the study on the visit days.

Pharmacokinetic parameters. Single and multiple dose tildrakizumab plasma concentrations and derived pharmacokinetic parameters including half-life, T_{max} (the time when C_{max} is observed), C_{max} and AUC were assessed. For part 2 only, preliminary single and multiple dose proportionality was assessed using a one-way ANOVA model extracting the effect due to dose. Ratio estimate and 90% confidence intervals were provided for the higher dose versus the lower dose in part 2.

Pharmacodynamic parameters. The primary efficacy end point was the percentage change in PASI score from baseline at 4 or 8 weeks after the last treatment day (that is, week 16 for parts 1, 2 and 3). PASI is a commonly used index for the evaluation of pharmacologic agents in the treatment of psoriasis[27,28].

Histology, immunohistochemistry and gene expression of skin biopsies. Four-millimetre punch biopsies were obtained in duplicate from lesional and non-lesional skin. For reasons of comparability, target sites of biopsy specimens were always located on trunk, thighs or arms for microscopic evaluations. Skin samples were either fixed in formalin and paraffin embedded or embedded in optimum cutting tissue compound (Tissue-Tek, Sakura Finetek Europe B.V.), snap-frozen in liquid nitrogen, and stored at $-80\,^{\circ}C$ until further processing. Samples for histology were cut from formalin-fixed paraffin-embedded blocks, stained with haematoxylin and eosin, and analysed by light microscopy. Pre-dose and post-dose biopsies for histological and immunohistochemical analysis were analysed from five placebo-treated patients, ten patients treated with 3 mg kg^{-1} tildrakizumab, and seven patients treated with 10 mg kg^{-1}. Immunohistochemical stainings of the study patients were compared to stainings of normal human skin samples from a previous study[29].

Histopathologic psoriasis severity score. Criteria used to calculate the HPSS were chosen on the basis of well-established histopathological criteria for the diagnosis of plaque-type psoriasis[30]. Specifically, changes affecting the epidermis, involvement of the blood vascular component and the amount and distribution of the inflammatory cell infiltrate are assessed. The changes are graded according to severity into 0 (absent), 1 (mild), 2 (moderate) and 3 (severe), respectively. The individual scores for each main category (epidermal, vascular, inflammatory cells) are added to obtain a total HPSS. The HPSS can range from 0 to 36, while each subcategory can yield a score between 0 and 12, depending on the grading (0–3) of the individual

features of each category. Extended Data Table 5 highlights the individual subcategories that sum up to each of the main categories. Histopathological analysis was performed independently by two dermatopathologists in a blinded fashion and Bland–Altman plots were used to assess the agreement between readings of the HPSS of two independent observers. Bland–Altman analysis showed good agreement between HPSS readings of two independent readers owing to the mean difference between readings of observer 1 and observer 2 being −0.3 (95% limits of agreement: −2.8 to 2.5) and thus not significantly different from zero.

Immunohistochemical staining. Immunohistochemistry was performed on frozen tissue as described before[29]. The purified monoclonal antibodies used for these stainings are shown in Extended Data Table 6. In brief, after cutting, mounting and fixing the tissue on slides, the primary antibodies were applied overnight in a humid chamber at 4 °C. Using a staining kit (Vectastain Elite ABC kit, mouse IgG, Vector Laboratories), sections were then incubated with a matching secondary antibody followed by incubation with an avidin-biotin complex for signal amplification. Finally, cells were visualized using 3-amino-9-ethyl-carbazol (Sigma-Aldrich) and counterstained with haematoxylin (Merck).

Immunofluorescence staining. To determine the phenotype and number of infiltrating leukocyte subsets, multicolour immunofluorescence stainings were performed as described before[31]. The monoclonal antibodies used and their sources are shown in Extended Data Table 6.

In brief, after incubation of the unconjugated primary antibody overnight at 4 °C, slides were washed and an appropriate second step was applied (Rhodamine (TRITC)-conjugated AffiniPureF(ab')$_2$ fragment goat anti-mouse IgG (H+L), Jackson ImmunoResearch Laboratories, Northern Light 637-conjugated donkey anti-mouse IgG, R&D Systems or Alexa 633-conjugated donkey anti-goat IgG, Invitrogen, Molecular Probes). After blocking with normal mouse serum (DAKO) to diminish background problems, secondary antibodies, either biotinylated or fluorescence-labelled, were applied for at least 6 h at 4 °C. Sections were further incubated with streptavidin TRITC (Jackson ImmunoResearch Laboratories), Streptavidin Alexa 488 (Invitrogen, Molecular Probes) or Streptavidin 637 NL (R&D Systems) and/or Oregon Green-labelled goat IgG anti-FITC (Invitrogen, Molecular Probes) to visualize antibody binding. Finally, nuclei were visualized by incubation with 4′,6-diamidino-2-phenylindole (DAPI).

In all staining experiments, substitution of the antibodies with isotype-matched immunoglobulin (purified, biotinylated or fluorescence-labelled) served as negative controls.

Evaluation of immunohistochemical and immunofluorescence results. Data acquisition was performed by using the TissueFAXS technology (TissueGnostics), based on a motorized Zeiss Observer Z1 automated multichannel fluorescence microscope, as previously described[32].

HistoQuest and TissueQuest softwares (TissueGnostics) were applied for image cytometry of immunohistochemical and immunofluorescence stainings, respectively. Positive cells in the epidermal and dermal compartments were analysed separately. For measuring epidermal thickness, the epidermis was randomly marked at ten different sites. Thickness was then determined by calculating the mean value of these ten measurements. For evaluation of keratin 16, a score ranging from 0 (healthy skin) to 4 (maximum psoriatic lesion) was used[33].

Confocal immunofluorescence microscopy (LSM 510, Zeiss) was used to further document immunofluorescence images at a higher magnification.

Gene expression analysis. Whole-cell lysates were prepared from slides of formalin-fixed paraffin-embedded tissue for analysis on the NanoString nCounter gene expression platform (NanoString Technologies). Before making the cell lysate, tissue sections were deparaffinized in xylene for 3 × 5 min and then rehydrated by immersing consecutively in 100% ethanol for 2 × 2 min, 95% ethanol for 2 min, 70% ethanol for 2 min and then immersed in dH_2O until ready to be processed. Tissue was lysed on the slide by adding 10–50 μl of PKD buffer. Tissue was scraped from the slide and transferred to a 1.5 ml Eppendorf tube. Proteinase K was added at no more than 10% final volume and the RNA lysate was incubated for 15 min at 55 °C and then 15 min at 80 °C. The RNA lysate was stored at −80 °C until gene expression profiling was performed using the NanoString nCounter system on the following genes: *CAMP*, *CCL20*, *CCR6*, *CXCL8*, *DEFB104A* (also known as *DEFB4*), *IFNG*, *IL12B*, *IL17A*, *IL17C*, *IL17F*, *IL19*, *IL20*, *IL22*, *IL23p19* (also known as *IL23A*), *IL23R*, *IL24*, *LCN2*, *RORC* and *S100A7*.

Statistical methods. Subjects were randomized in this study to balance the preconditions of the subjects in the different treatment groups. This study was designed to enroll approximately 64 subjects (24 in part 1, and 40 in part 2). Additional subjects were to be enrolled at the discretion of the study sponsor to explore further interim doses or to increase the power of the statistical analyses. The sample size for parts 1 and 3 was determined by practical considerations only as it was designed to assess safety. In part 2, subjects were assigned to each of the dose levels in a 3:1 ratio (that is, 15 subjects on active and 5 on placebo in each of the dose groups). No statistical methods were used to predetermine sample size.

With 15 subjects on tildrakizumab and 16 (6 + 5 + 5) subjects on placebo, the study was designed to detect a difference of approximately 30% in the mean change from baseline PASI score compared with the placebo group. This assumes a 0% mean change from baseline in the placebo group and a standard deviation of 40%, respectively, with 80% power at an alpha level of 0.1 (one sided test). Subjects were pre-planned to be pooled for the purpose of increasing the power of the statistics test.

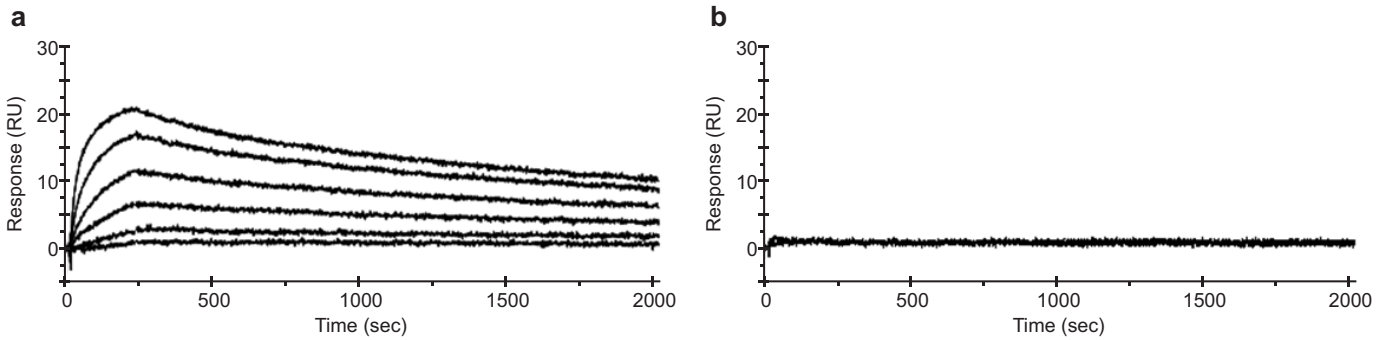
The percent change in PASI score was analysed using a one-way ANOVA model extracting the effect due to treatment. The mean and 95% confidence interval was provided for each treatment group as well as the difference between tildrakizumab and placebo. Durability of effect was evaluated in an exploratory and graphical fashion.

For HPSS analysis, statistical comparison between psoriasis patients before and after treatment with tildrakizumab were performed using the non-parametric Wilcoxon matched-pairs signed-rank test calculated using the GraphPad Prism Software. Bland–Altman plots were used to assess the agreement between readings of the HPSS of two independent observers. The data are presented as the mean and s.e.m.

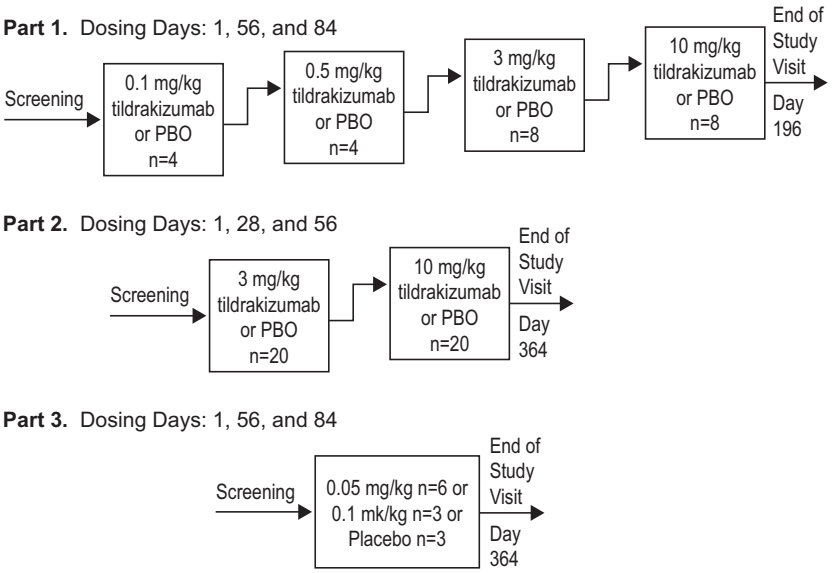
For immunohistological and immunofluorescence analysis, statistical comparison between psoriasis patients before treatment with tildrakizumab and after treatment were performed using the non-parametric Wilcoxon matched-pairs signed-rank test calculated using the GraphPad Prism Software. The data are presented as the mean and s.e.m.

For gene expression analysis, statistical comparison between psoriasis patients before treatment with tildrakizumab and after treatment were performed using the non-parametric Wilcoxon matched-pairs signed rank test calculated using the GraphPad Prism Software.

27. Fredriksson, T. & Pettersson, U. Severe psoriasiseoral therapy with a new retinoid. *Dermatologica* **157,** 238–244 (1978).
28. Langley, R. G. Effective and sustainable biologic treatment of psoriasis: what can we learn from new clinical data? *J. Eur. Acad. Dermatol. Venereol.* **26** (suppl. 2), 21–29 (2012).
29. Stary, G., Bangert, C., Stingl, G. & Kopp, T. Dendritic cells in atopic dermatitis: expression of FcepsilonRI on two distinct inflammation-associated subsets. *Int. Arch. Allergy Immunol.* **138,** 278–290 (2005).
30. Ackerman, A. B., Boer, A., Bennin, B. & Gottlieb, G. J. *Histologic Diagnosis of Inflammatory Skin Diseases: an Algorithmic Method Based on Pattern Analysis* 3rd edn (Ardor Scribendi, 2005).
31. Bangert, C., Friedl, J., Stary, G., Stingl, G. & Kopp, T. Immunopathologic features of allergic contact dermatitis in humans: participation of plasmacytoid dendritic cells in the pathogenesis of the disease? *J. Invest. Dermatol.* **121,** 1409–1418 (2003).
32. Bangert, C. *et al.* Clinical and cytological effects of pimecrolimus cream 1% after resolution of active atopic dermatitis lesions by topical corticosteroids: a randomized controlled trial. *Dermatology* **222,** 36–48 (2011).
33. Skvara, H. *et al.* The PKC inhibitor AEB071 may be a therapeutic option for psoriasis. *J. Clin. Invest.* **118,** 3151–3159 (2008).



Extended Data Figure 1 | Tildrakizumab binds to IL-23, but not IL-12. **a**, **b**, Biacore sensorgrams showing the 37 °C binding kinetics of human IL-23 (**a**) and human IL-12 (**b**) to immobilized tildrakizumab as a threefold dilution series from 0.091 to 22.2 nM. The association phase was measured for 220 s followed by a dissociation phase measurement for 1,800 s. The affinity of tildrakizumab for IL-23 is ~300 pM.



Extended Data Figure 2 | Study flow chart.

Extended Data Table 1 | Baseline patient demographics

	0.05 mg/kg (Part 3) n=6	0.1 mg/kg (Part 1 & 3) n=6	0.5 mg/kg (Part 1) n=3	3 mg/kg (Part 1)* n=7	3 mg/kg (Part 2) n=15	10 mg/kg (Part 1) n=6	10 mg/kg (Part 2) n=14	Placebo (Parts 1, 2, & 3) n=20	Total n=77
Sex (n,%)									
Female	2 (33)	1 (17)	1 (33)	0	5 (33)	2 (33)	1 (7)	4 (20)	16 (21)
Male	4 (67)	5 (83)	2 (67)	7 (100)	10 (67)	4 (67)	13 (93)	16 (80)	61 (79)
Race (n,%)									
White	5 (83)	6 (100)	2 (67)	5 (71)	15 (100)	4 (67)	13 (93)	19 (95)	69 (90)
American Indian or Alaskan Native	0	0	1 (33)	1 (14)	0	0	0	0	2 (3)
Asian	1 (17)	0	0	1 (14)	0	0	1 (7)	0	3 (4)
Black or African American	0	0	0	0	0	2 (33)	0	1 (5)	3 (4)
Age (yrs)									
Mean (SD)	45.0 (9.1)	48.2 (11.7)	47.3 (11.9)	52.7 (6.2)	49.3 (11.7)	46.2 (11.1)	46.0 (14.4)	45.5 (11.6)	47.3 (11.4)
Range	33 - 55	31 - 62	34 - 57	42 - 61	22 - 63	32 - 65	22 - 64	25 - 61	22 - 65
Age (n,%)									
18 - <65	6 (100)	6 (100)	3 (100)	7 (100)	15 (100)	5 (83)	14 (100)	20 (100)	76 (99)
65 or Older	0	0	0	0	0	1 (17)	0	0	1 (1)
Weight (kg)									
Mean (SD)	96.88 (20.37)	95.85 (23.49)	107.63 (24.58)	90.27 (10.96)	96.51 (10.87)	94.25 (25.88)	95.69 (20.46)	102.46 (25.49)	97.67 (20.23)

* Includes one subject that was randomized but never treated.

Extended Data Table 2 | Subjects on tildrakizumab who achieved PASI75 and PASI90 on day 112

	Parts 1 and 3 Pooled					Part 2	
Tildrakizumab dose	0.05 mg/kg	0.1 mg/kg	0.5 mg/kg	3 mg/kg	10 mg/kg	3 mg/kg	10 mg/kg
	n/N(%)	n/N(%)	n/N(%)	n/N(%)	n/N(%)	n/N(%)	n/N(%)
PASI 75	2/5 (40%)	3/6 (50%)	1/3 (33%)	6/6 (100%)	3/5 (60%)	10/15 (66%)	13/14 (93%)
PASI 90	2/5 (40%)	1/6 (17%)	0/3 (0%)	5/6 (83%)	1/5 (20%)	7/15 (47%)	10/14 (71%)

Extended Data Table 3 | Summary of adverse events occurring in 2 or more subjects within a treatment group

Adverse Event	0.05 mg/Kg (n=6) Part 3	0.1 mg/kg (n=6) Parts 1 and 3	0.5 mg/kg (n=3) Part 1	3 mg/kg (n=7) Part 1	3.0 mg/kg (n=15) Part 2	10 mg/kg (n=6) Part 1	10 mg/kg (n=14) Part 2	Placebo (n=20) ALL Parts	Total (n=77)
No Subjects reporting any adverse event (%)	6 (100)	4 (67)	3 (100)	5 (71)	15 (100)	2 (33)	12 (86)	15 (75)	62 (81)
Headache	1 (17)	1 (17)	1(33)	1 (14)	4 (27)	0	3 (21)	3 (15)	14 (18)
Upper respiratory tract infection	1 (17)	2 (33)	1 (33)	0	2 (13)	0	5(36)	3 (15)	14 (18)
Cough	0	1 (17)	0	1 (14)	4 (27)	0	3 (21)	3 (15)	12(16)
Nasopharyngitis	0	1 (17)	0	1(14)	4 (27)	0	4 (29)	2 (10)	12 (16)
Arthralgia	0	0	0	1 (14)	1 (7)	0	1 (7)	3 (15)	6 (8)
Back pain	0	0	0	1 (14)	2 (13)	0	0	1 (5)	4 (5)
Blood creatine phosphokinase increased	0	0	0	0	2 (13)	0	2 (14)	0	4 (5)
Hypertension	2 (33)	0	0	1(14)	0	0	0	1 (5)	4 (5)
Oropharyngeal pain	0	0	0	0	2 (13)	0	1 (7)	1 (5)	4 (5)
Rhinitis	1(17)	0	0	0	2 (13)	0	1 (7)	0	4 (5)
Eczema	0	0	0	0	2 (13)	0	1 (7)	0	3 (4)
Fatigue	0	0	0	0	1 (7)	0	0	2 (10)	3 (4)
Pruritus	0	0	0	0	2 (13)	0	0	1 (5)	3 (4)
Sinusitis	0	0	0	0	0	0	0	3 (15)	3 (4)
Insomnia	0	2 (33)	0	0	0	0	0	0	2 (3)
Leukopenia	0	0	0	0	2 (13)	0	0	0	2 (3)
Paraesthesia	0	0	0	0	2 (13)	0	0	0	2 (3)
Psoriasis	0	0	0	0	0	0	0	2 (10)	2 (3)
Sciatica	0	0	0	0	2 (13)	0	0	0	2 (3)

Extended Data Table 4 | Summary of serious adverse events

Subject	SAE	Treatment	Relationship to treatment
1	Upper limb fracture	0.05 mg/kg tildrakizumab	Unrelated
2	Ankle fracture	0.05 mg/kg tildrakizumab	Unrelated
3	Pneumothorax	3 mg/kg tildrakizumab	Unrelated
4	Meniscal abrasion	3 mg/kg tildrakizumab	Unrelated
5*	Cardiac enzymes increased Myositis Cytomegalovirus infection	3 mg/kg tildrakizumab	Unrelated
6	Convulsion	10 mg/kg tildrakizumab	Possibly related**
7	Pulmonary embolism Pulmonary hypertension	10 mg/kg tildrakizumab	Unrelated
8	Facial bones fracture	10 mg/kg tildrakizumab	Unrelated

* Subject 5 had a serious adverse event (SAE) of increased cardiac enzymes (44 weeks after treatment) and myositis (58 weeks after treatment); treatment of the condition with prednisone likely contributed to the cytomegalovirus infection (59 weeks after treatment). The primary investigator did not consider these serious adverse events to be related to study medication. **In subject 6, the convulsion event was confounded by the patient's recent history of alcohol abuse, acute alcohol and benzodiazepine withdrawal as well as lack of sleep preceding the event. The investigator considered the event to possibly be related to the study drug owing to its long half-life; the event occurred 17 days after the second dose of 10 mg kg^{-1} tildrakizumab.

Extended Data Table 5 | The histopathologic psoriasis severity score

Main Categories	Subcategories (grading 0-3 each)				Score
Epidermal changes	Hyperkeratosis	Hyperplasia	Hypogranulosis	Mitosis	0-12
Vascular changes	Dilated vessels	Elongated vessels	Increased number	Contorted vessels	0-12
Inflammatory cells	Perivascular lymphocytes	Neutrophils intraepidermal	Neutrophils intracorneal	Histiocytes intradermal	0-12
Total HPSS					0-36

Extended Data Table 6 | Antibodies used for immunohistochemistry and immunofluorescence

Antigen	Supplier	Clone	Isotype	Catalog Number	Reference
IL23p19 pur.	Merck	12F12	mIgG2a	-	Wilson NJ et al 2007, Nat Immunol. 8(9):950-7
CD31 pur.	Dako	JC70A	mIgG1	M0823	Parums DV et al 1990, J Clin Pathol 43:572-577
Ki-67 pur.	Dako	MIB-1	mIgG1	MIB-1	Gerdes J et al 1992, J Pathol; 168:85-6
Keratin 16 pur.	Neomarkers	LL025	mIgG1	MS-620-PO	Wetzels RH et al 1991, Am J Pathol; 138:751-763
CD3 pur.	Beckman Coulter	UCHT1	mIgG1	IM1304	Thibault G et al 1995, J Immunol; 154: 3814-3820
CD8 pur.	Dako	C8/144B	mIgG1	M7103	Mason DY et al 1992, J Clin Pathol; 45: 1084-88
CD4 pur.	Dako	MT310	mIgG1	M0716	Leong AS-Y et al 1999; London: Oxford University Press: 49-50
CD11c pur.	IOTest/Beckmann Coulter	BU15	mIgG1	IM0712	Kohrgruber N et al J. Immunol; 163: 3250-3259
BDCA-2 pur.	Dendritics	104C12.08	mIgG1	DDX0041	Stary G et al 2009, Blood; 114: 3854-3863
Langerin pur.	IOTest/Immunotech	DCGM4	mIgG1	IM3449	Valladeau J et al 2000, Immunity; 12: 71-81
CD117 APC	Immunotech	104D2D1	mIgG1	IM3638	Uoshima N et al 1995, Br. J. Haematol; 91: 30-36
CD11c biot	Biolegend	3.9	mIgG1	301612	Gurer C *et al* 2008, *Blood*; 112:1231-1239
CD15 biot	Novus Biologicals	HI98	mIgM	NBP1-43673	Lund-Johansen F et al 1992, J Immunol. 148: 3221-3229 *
CD163 FITC	Acris	5C6-FAT	mIgG1	BM4041F	Zaba LC et al 2007, J Exp Med; 204:3183-3194

* Novus Biologicals was only founded in summer 1996. Publication cited refers to the same clone (HI98) from a different company.

LETTER

doi:10.1038/nature14226

Nuclear architecture dictates HIV-1 integration site selection

Bruna Marini[1], Attila Kertesz-Farkas[2]*, Hashim Ali[1]*, Bojana Lucic[1]†, Kamil Lisek[1]†, Lara Manganaro[1]†, Sandor Pongor[2]†, Roberto Luzzati[3,4], Alessandra Recchia[5], Fulvio Mavilio[5,6], Mauro Giacca[1,4]§ & Marina Lusic[1]§†

Long-standing evidence indicates that human immunodeficiency virus type 1 (HIV-1) preferentially integrates into a subset of transcriptionally active genes of the host cell genome[1–4]. However, the reason why the virus selects only certain genes among all transcriptionally active regions in a target cell remains largely unknown. Here we show that HIV-1 integration occurs in the outer shell of the nucleus in close correspondence with the nuclear pore. This region contains a series of cellular genes, which are preferentially targeted by the virus, and characterized by the presence of active transcription chromatin marks before viral infection. In contrast, the virus strongly disfavours the heterochromatic regions in the nuclear lamin-associated domains[5] and other transcriptionally active regions located centrally in the nucleus. Functional viral integrase and the presence of the cellular Nup153 and LEDGF/p75 integration cofactors are indispensable for the peripheral integration of the virus. Once integrated at the nuclear pore, the HIV-1 DNA makes contact with various nucleoporins; this association takes part in the transcriptional regulation of the viral genome. These results indicate that nuclear topography is an essential determinant of the HIV-1 life cycle.

One important aspect of the interaction between HIV-1 and its target cells is the encounter between the viral complementary DNA (cDNA) with the complex architecture of the mammalian nucleus, in which chromosomes and genes are spatially arranged to occupy preferred positions within the three-dimensional space[6].

We analysed the lists of human genes targeted by HIV-1 from six different studies (Extended Data Table 1), containing altogether 1,136 unique gene integration sites in activated T cells carrying the CD4 antigen ($CD4^+$); 126 of these genes recurred in two lists, 24 in three, and six in at least four lists, for a total of 156 genes, which we named HIV recurrent integration genes (RIGs). The probability of detecting this number of specific genes by chance was extremely low ($P < 1 \times 10^{-9}$; Extended Data Fig. 1a). RIGs were also highly represented in another list of approximately 12,000 integration sites[4], 5,221 of which were unique genes, as well as in two integration lists generated from patients' $CD4^+$ T cells[7,8] ($P < 0.001$ of detecting these genes by chance). Thus, RIGs are bona fine the hottest spots of HIV-1 integration.

We then ranked RIGs according to their frequency and plotted them onto the human chromosome map[9]. Unexpectedly, they appeared to cluster into specific chromosomal regions (Extended Data Fig. 1b). In five out of eight cases, RIGs were also in proximity to the 'hotter zones', previously defined as regions with remarkably high HIV-1 integration density[1] (Supplementary Table 1). In these areas, observations hinted at the possibility that the topological distribution of these chromosomal regions inside the nucleus could determine HIV-1 integration.

By applying three-dimensional immuno-DNA fluorescence *in situ* hybridization (FISH), we assessed the position of RIGs and hotter zones in primary $CD4^+$ T cells from healthy donors. Selected FISH probes, listed in the Supplementary Information, provided topological information for a total of 169 RIGs and other integration sites located within 10 megabases (Mb) from the centre of the probe (Extended Data Fig. 2).

When the radial positions of the RIG FISH signals were binned into three zones of equal area[10,11] (Fig. 1a), a clear gradient in signal localization was observed, which decreased from the nuclear envelope towards the interior (images of 14 RIGs in Fig. 1b, c; four hotter zones in Fig. 1d). The global distribution of RIGs ($n = 1{,}420$ analysed alleles) was remarkably different from that of control genes, all of which were expressed in $CD4^+$ T cells[12,13] ($n = 522$): 44% of RIGs mapped in zone 1, 41.5% in zone 2 and only 14.5% in zone 3 versus 25.6%, 47.6% and 26.8% for control genes, respectively (Fig. 1f; representative images of control genes are shown in Fig. 1e). Considering an average of about 7 μm for the nuclear diameter in $CD4^+$ T cells, 63% of RIGs and hotter-zone alleles were concentrated within about 1 μm below the nuclear membrane.

We wanted, therefore, to visualize the position of the HIV-1 DNA itself in infected, primary $CD4^+$ T-cell nuclei. At 4 days after infection with the VSV-G-pseudotyped HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$[14], the vast majority of the proviral immuno-FISH signals were in zone 1 (75.2% within 1 μm under the nuclear envelope) (Fig. 2c). The visualized viral DNA was integrated[15], as also detected by real-time Alu PCR (Fig. 2a), and transcriptionally active (Fig. 2b). A similar distribution was observed in primary macrophages and the monocytic cell line U937 (Extended Data Fig. 3a, b, respectively). Peripheral localization was also observed for a fully competent virus carrying the HIV-1_{BRU} envelope[16] (Fig. 2d) and, notably, for the wild-type viruses found in $CD4^+$ T-cells from two HIV-infected patients (Fig. 2e, f). Peripheral localization was also a feature of lentiviral vectors, irrespective of their transcriptional activity (Extended Data Fig. 3c, d), but not of the MoMLV gammaretrovirus, which localized preferentially inside the nuclear interior (Extended Data Fig. 3e).

In contrast, when integration was impaired, the viral cDNA roamed around the nucleus. This was the case for two HIV-1 clones harbouring single-point mutations in the integrase catalytic domain (class I IN mutations: IN(D64E) and IN(D116N))[17,18] or for HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$ in the presence of the integrase inhibitor raltegravir; under these conditions, only 10–20% of viruses were found in zone 1 (Fig. 2g). In these cases, the detected viral genomes did not correspond to integrated DNA (Fig. 2h) but were highly enriched in circular forms of viral DNA containing two long terminal repeats (2-LTR circles) (Fig. 2i). We also downregulated the chromatin tethering factor LEDGF/p75 (ref. 19) and the inner nuclear basket protein Nup153 (ref. 20), which are involved in viral DNA integration (Fig. 2j). FISH was performed 48 h after infection when there

[1]Molecular Medicine Laboratory, International Centre for Genetic Engineering and Biotechnology (ICGEB), 34149 Trieste, Italy. [2]Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), 34149 Trieste, Italy. [3]Struttura Complessa Malattie Infettive, Azienda Ospedaliero-Universitaria, 34134 Trieste, Italy. [4]Department of Medical, Surgical and Health Sciences, University of Trieste, 34129 Trieste, Italy. [5]Department of Life Sciences, University of Modena and Reggio Emilia, 41121 Modena, Italy. [6]Genethon, 91002 Evry, France. †Present addresses: Department of Infectious Diseases, Integrative Virology, University Hospital Heidelberg and German Center for Infection Research, 69120 Heidelberg, Germany (B.L.; M.L.); Laboratorio Nazionale Consorzio Interuniversitario per le Biotecnologie (LNCIB), 34149 Trieste, Italy (K.L.); Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA (L.M.); Pazmany University, Budapest 1083, Hungary (S.P.).

*These authors contributed equally to this work.

§These authors jointly supervised this work.

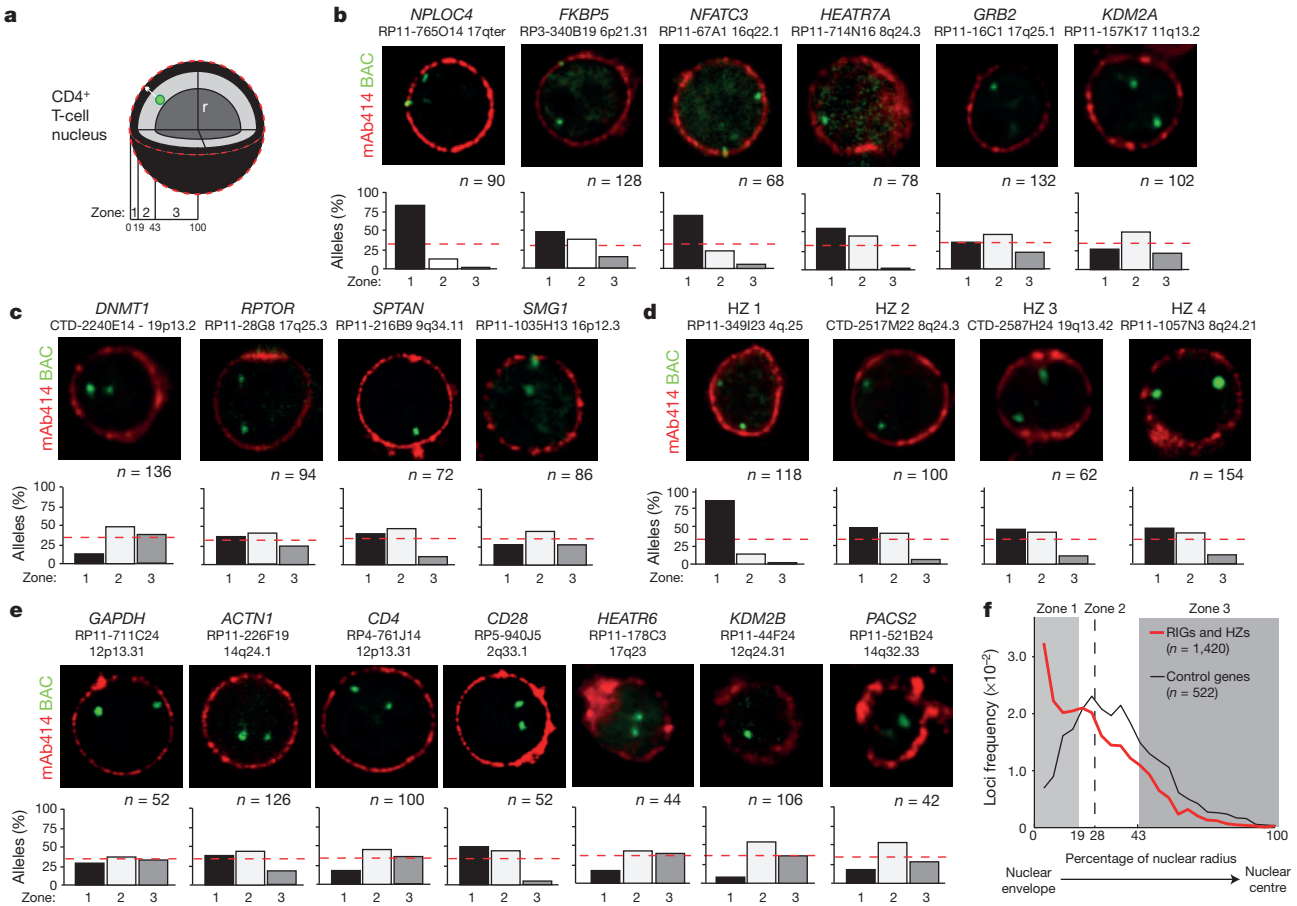


Figure 1 | Localization of HIV RIGs at the nuclear periphery. **a**, Subdivision of nucleus into three concentric zones of equal area. **b**–**e**, Three-dimensional immuno-DNA FISH of ten HIV RIGs (**b**, **c**), four hotter zones (**d**) and seven control genes (**e**) in activated CD4$^+$ T cells (green: bacterial artificial chromosome (BAC) probe labelled with DIG (dUTP-digoxygenin) and FITC (fluorescein isothiocyanate); red: NPC staining by mAb414). Below each representative image, the distribution of the analysed alleles into the three nuclear zones is shown, normalized over nuclear radius. Evenly distributed random genes would be enriched equally in the three zones (red dashed line). The number of alleles analysed is shown at the bottom of each panel. HZ, hotter zone. **f**, Distribution of the relative distances of all measured alleles from the nuclear envelope (HIV RIGs and hotter zones: $n = 1{,}420$; control genes: $n = 522$). The three zones are shown by grey shading. The dashed line indicates approximately 1 μm from the nuclear edge of the T-cell nucleus.

was a marked reduction in HIV-1 integration (Fig. 2k), and the majority of the signals labelled unintegrated viral DNA[16]: 72% and 77% of the FISH signals were in zones 2 and 3 for the LEDGF/p75 and Nup153 knockdowns, respectively (Fig. 2l). The effect of the Nup153 knockdown was rescued by transfecting an expression plasmid coding for an RNA interference (RNAi)-resistant Nup153 (Extended Data Fig. 4).

Most of the HIV-1 targets are common in different cell types; however, subtle differences exist. For example, HIV-1 almost never targets the *IKZF3* locus in CD34$^+$ haematopoietic stem cells[21] ($P < 1 \times 10^{-12}$), whereas the *TAP2* gene from the major histocompatibility complex class II locus is never targeted in CD4$^+$ T cells ($P < 1 \times 10^{-13}$). Strikingly, we observed that *IKZF3* localized in zones 1 and 2 in peripheral blood CD4$^+$ T cells (>80% of alleles), while it was almost absent from zone 1 in cord blood CD34$^+$ cells (<6% alleles; $P < 0.001$). Conversely, the *TAP2* locus was absent from zone 1 in CD4$^+$ T cells (<8% of alleles), while it was distributed between zones 1 and 2 in CD34$^+$ cells (>90% of alleles; $P < 0.001$; Fig. 3a).

To understand the chromatin features of RIGs, we compared the available data from chromatin immunoprecipitation sequencing (ChIP-seq) obtained in CD4$^+$ T cells for RIGs[22], cold genes (defined as transcriptionally inactive genes never targeted by HIV-1; F.M. and A.R., unpublished observations) and a list of genes corresponding to the 1,000 most expressed (active) and 1,000 least expressed (silent) genes from the GNF SymAtlas[13]. Association of RNA Pol2 with RIGs had a pattern superimposable to that of active genes, peaking at the transcription start sites (TSSs) (Fig. 3b). In a similar manner, distribution of markers of active transcription (H3K9ac, H3K36me3, H3K4me3, H4K16ac and H4K20me) was identical for RIGs and active genes (Fig. 3c–e and Extended Data Fig. 5). In contrast, markers of facultative (H3K9me2) and constitutive (H3K9me3 and H3K27me3) chromatin were found enriched both on cold genes (where HIV-1 never integrates) and on silent genes, but not on RIGs (Extended Data Fig. 5). Of interest, active genes and RIGs had a superimposable distribution of H3K4me2, which is enriched at the lamin-associated domain (LAD) borders[5] (Fig. 3f).

Heterochromatic LADs contain approximately 4,000 transcriptionally inactive genes[5]. We found that more than 90% of HIV RIGs lay outside LADs, while almost 80% of cold genes were inside LADs ($P < 0.001$ compared with a random gene distribution; Fig. 3g). Immuno-FISH images for three of these cold genes confirmed their localization close to the nuclear envelope in primary CD4$^+$ T cells (Fig. 3h). Finally, when all the 1,344 known LADs were aligned by their left or right borders, 87.2% of RIGs were found outside the LADs, in contrast to a random distribution of genes (68.2%; $P < 0.001$, also taking into account the lower gene density within LADs; Fig. 3i).

Transcriptionally active genes at the nuclear periphery are often associated with the nuclear pore complex (NPC)[23–27]. We therefore assessed interaction of the HIV-1 provirus with the NPC by ChIP assays in primary CD4$^+$ T cells (primer scheme and controls in Extended Data Fig. 6a, b). At 4 days after infection, when RNA Pol2 and the USF1 and p65/RelA transcription factors were associated with the viral DNA as expected[14], both the mAb414 antibody, which recognizes phenylalanine–glycine (FG)-repeats in nucleoporins, and specific antibodies against Nup153,

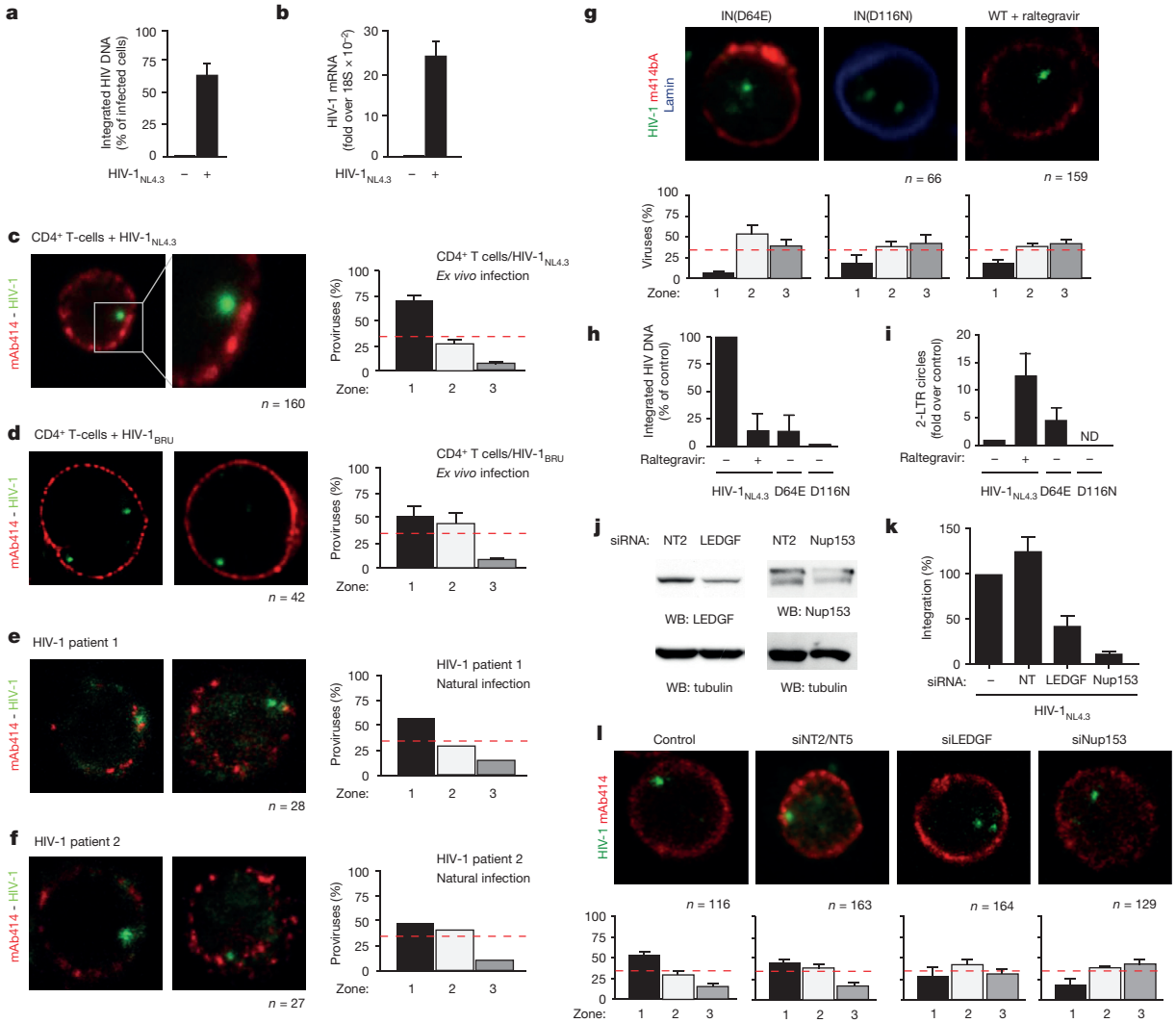


Figure 2 | Integrated, transcriptionally active HIV-1 is found at the nuclear periphery. **a**, **b**, Quantification of integrated HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$ DNA (**a**) and HIV RNA (**b**) by real-time Alu PCR in infected $CD4^+$ T cells. **c**–**f**, Three-dimensional immuno-DNA FISH of HIV-1 DNA (green) in primary $CD4^+$ T cells infected *ex vivo* with HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$ (**c**) and HIV-1_{BRU} (**d**), or directly obtained from two patients infected with HIV-1 (**e**, **f**). **g**, Three-dimensional immuno-DNA FISH of HIV-1 DNA in activated $CD4^+$ T cells infected with the mutant viruses IN(D64E) or IN(D116N) or with HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$ upon raltegravir treatment. **h**, **i**, Real-time Alu PCR (**h**) and 2-LTR quantification (**i**) in the cells treated as in **g**. ND, not determined. **j**, Western blot (WB) showing protein levels for LEDGF/p75 and Nup153 at the moment of HIV-1 infection, 36 h after short interfering RNA (siRNA) transfection. NT2, non-targeting siRNA. **k**, Real-time Alu PCR in Jurkat cells infected with HIV-$1_{NL4.3}$ and previously transfected with a non-targeting siRNA (NT) or an siRNA targeting LEDGF/p75. Samples were normalized over control-infected cells. **l**, Three-dimensional immuno-DNA FISH for HIV-1 DNA visualization upon Jurkat cell treatment with the indicated siRNAs. NT2 and NT5 are two non-targeting siRNAs. All graphs, except those relative to patients' cells, show mean and s.e.m. of at least three independent experiments.

Nup98, Nup62 and Tpr all immunoprecipitated the HIV-1 DNA; binding was also observed for the *NPLOC4* RIG gene, but not for the LAD gene *PTPRD* (Extended Data Fig. 6c). When ChIP was performed on the IN-defective D64E virus, no viral DNA was detected using the mAb414 and anti-Nup153 antibodies (Extended Data Fig. 6d).

Next, we aimed to verify whether HIV-1 localization changed when the virus reverted from a transcriptionally inactive to an active state. In the latent T-cell J-Lat clone 15.4 (ref. 28), the HIV-1 DNA retained its gross peripheral localization both in inactive and in TPA (12-*O*-tetradecanoylphorbol-13-acetate) phorbol ester-reactivated conditions (Extended Data Fig. 7a, b). Similar results were obtained in a primary model of HIV-1 latency[14] (Extended Data Fig. 7c–e). However, when localization was analysed at molecular resolution by ChIP using the mAb414, anti-Tpr and anti-Nup153 antibodies, binding of the proviral region located downstream of the TSS to the nucleoporins was observed upon transcriptional activation but not in latent conditions (Extended Data Fig. 7f). We also observed that nucleoporins directly participated in HIV-1 transcriptional regulation. When Tpr and Nup153 were silenced by RNAi in latent J-Lat cells, proviral transcription was significantly reduced (Extended Data Fig. 7g, h). Similarly, downregulation of Tpr also blunted LTR-driven gene expression in HIV-1-infected HeLa cells (Extended Data Fig. 8a–e).

Our findings show that the cellular genes that are highly targeted by HIV-1 are distributed in a topologically non-random manner, being positioned within 1 μm from the nuclear edge; these genes are enriched in open chromatin marks, excluded from the LADs and associated with the NPC. Thus, the HIV-1 pre-integration complex preferentially targets those areas of open chromatin that are proximal to the nuclear pore, while excluding the internal regions in the nucleus as well as the peripheral regions associated with the nuclear lamina (model in Extended Data Fig. 9). The localization of HIV-1 proviral DNA in close association with the nuclear pore is consistent with several observations showing that different NPC components play a role in HIV-1 infection[20,29,30].

Why does the viral DNA integrate into the NPC compartment? A possibility that we favour is that the virus simply integrates into the first open chromatin regions it meets along its route into the nucleus. This is

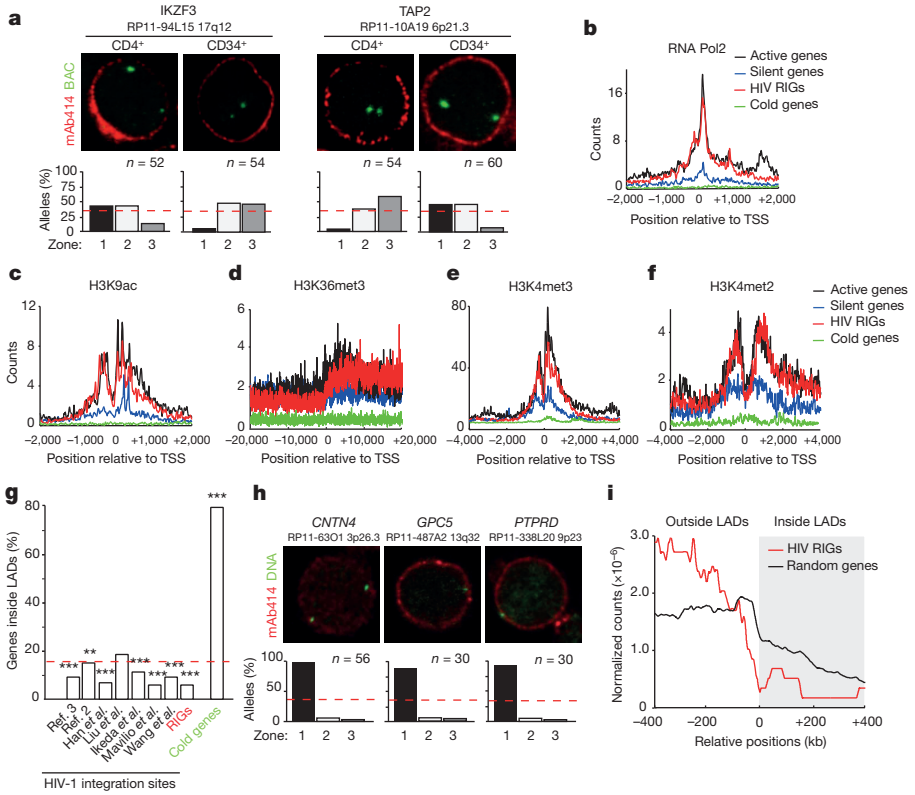


Figure 3 | HIV RIGs are transcriptionally active genes that are excluded from the LADs. **a**, Localization of the *IKZF3* (**a**) and *TAP2* (**b**) genes (green) in $CD4^+$ T cells and $CD34^+$ haematopoietic stem cells. **b**–**f**, Distributions of Pol2, acetylated H3K9, H3K36me3 and H3K4me2/3 around the TSSs of HIV RIGs (red) and cold genes (green), compared with highly active (black) and silent (blue) genes in activated $CD4^+$ T cells. **g**, Cross-comparison of different lists of integration loci, including HIV RIGs, with the lists of genes present inside LADs: HIV integration loci are significantly depleted in LADs compared with a null distribution (indicated by a red dotted line). $***P < 0.001$, $**P < 0.01$, $*P < 0.05$. References with authors' names can be found in the reference list at the end of the Methods section. **h**, Three-dimensional immuno-DNA FISH in activated $CD4^+$ T cells of three cold genes predicted to be inside LADs by bioinformatics analysis. **i**, Distribution of HIV RIGs (red) and of a random set of genes (black) around aligned LAD border regions. The light grey area with positive genomic coordinates indicates the regions inside LADs; the white area with negative coordinates is outside LADs.

likely to be related to the short life of viral integrase[16] and thus the need, for the pre-integration complex, to achieve rapid integration into genomic DNA upon its entry into the nucleus. This interpretation is consistent with our observation of more dispersed, unintegrated viral cDNA in all conditions in which integrase function is impaired.

Finally, while adding a three-dimensional view to the process of HIV-1 integration, our results also indicate that the localization of the HIV-1 DNA in close correspondence with the nuclear pore has functional relevance, since it appears important for productive HIV-1 gene expression.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 December 2013; accepted 9 January 2015.
Published online 2 March 2015.

1. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17,** 1186–1194 (2007).
2. Schroder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110,** 521–529 (2002).
3. Brady, T. *et al.* HIV integration site distributions in resting and activated $CD4^+$ T cells infected in culture. *AIDS* **23,** 1461–1471 (2009).
4. Sherrill-Mix, S. *et al.* HIV latency and integration site placement in five cell-based models. *Retrovirology* **10,** 90 (2013).
5. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
6. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nature Struct. Mol. Biol.* **20,** 290–299 (2013).
7. Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–183 (2014).
8. Wagner, T. A. *et al.* Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345,** 570–573 (2014).
9. Kin, T. & Ono, Y. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* **23,** 2945–2946 (2007).
10. Nagai, S. *et al.* Functional targeting of DNA damage to a nuclear pore-associated SUMO-dependent ubiquitin ligase. *Science* **322,** 597–602 (2008).
11. Hediger, F., Neumann, F. R., Van Houwe, G., Dubrana, K. & Gasser, S. M. Live imaging of telomeres: yKu and Sir proteins define redundant telomere-anchoring pathways in yeast. *Curr. Biol.* **12,** 2076–2089 (2002).
12. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10,** R130 (2009).
13. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
14. Lusic, M. *et al.* Proximity to PML nuclear bodies regulates HIV-1 latency in $CD4^+$ T cells. *Cell Host Microbe* **13,** 665–677 (2013).
15. Butler, S. L., Hansen, M. S. & Bushman, F. D. A quantitative assay for HIV DNA integration *in vivo*. *Nature Med.* **7,** 631–634 (2001).
16. Manganaro, L. *et al.* Concerted action of cellular JNK and Pin1 restricts HIV-1 genome integration to activated $CD4^+$ T lymphocytes. *Nature Med.* **16,** 329–333 (2010).
17. Lu, R., Limon, A., Ghory, H. Z. & Engelman, A. Genetic analyses of DNA-binding mutants in the catalytic core domain of human immunodeficiency virus type 1 integrase. *J. Virol.* **79,** 2493–2505 (2005).
18. Negri, D. R. *et al.* Successful immunization with a single injection of non-integrating lentiviral vector. *Mol. Ther.* **15,** 1716–1723 (2007).
19. Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21,** 1767–1778 (2007).
20. Matreyek, K. A., Yucel, S. S., Li, X. & Engelman, A. Nucleoporin NUP153 phenylalanine-glycine motifs engage a common binding pocket within the HIV-1 capsid protein to mediate lentiviral infectivity. *PLoS Pathog.* **9,** e1003693 (2013).
21. Cattoglio, C. *et al.* High-definition mapping of retroviral integration sites defines the fate of allogeneic T cells after donor lymphocyte infusion. *PLoS ONE* **5,** e15688 (2010).
22. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
23. Capelson, M. *et al.* Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. *Cell* **140,** 372–383 (2010).

24. Kalverda, B., Pickersgill, H., Shloma, V. V. & Fornerod, M. Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell* **140,** 360–371 (2010).
25. Vaquerizas, J. M. *et al.* Nuclear pore proteins nup153 and megator define transcriptionally active regions in the *Drosophila* genome. *PLoS Genet.* **6,** e1000846 (2010).
26. Liang, Y., Franks, T. M., Marchetto, M. C., Gage, F. H. & Hetzer, M. W. Dynamic association of NUP98 with the human genome. *PLoS Genet.* **9,** e1003308 (2013).
27. Light, W. H. *et al.* A conserved role for human Nup98 in altering chromatin structure and promoting epigenetic transcriptional memory. *PLoS Biol.* **11,** e1001524 (2013).
28. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells *in vitro*. *EMBO J.* **22,** 1868–1877 (2003).
29. Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319,** 921–926 (2008).
30. Konig, R. *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* **135,** 49–60 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the Italian National Research Programme on AIDS of the Istituto Superiore di Sanità, Italy, to M.G. and M.L. and from the Young Investigator Grant RF2007-16 of the Italian Ministry of Health to M.L. The authors are grateful to S. Kerbavcic for editorial assistance.

Author Contributions B.M., B.L., K.L. and M.L. performed the immuno-DNA FISH and ChIP experiments; A.K.-F., B.M., S.P., M.L. and M.G. analysed the data; A.K.-F., B.M. and S.P. performed the bioinformatics analysis; H.A. and M.L. performed the experiments using infectious virus; L.M. generated and analysed integrase-defective HIV-1 molecular clones; R.L. contributed to studies in primary cells from patients with HIV; A.R. and F.M. generated lentiviral vectors and analysed integration into $CD4^+$ T cells and $CD34^+$ bone marrow cells; M.L. and M.G. conceived and supervised the experiments and wrote the paper with help from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G. (giacca@icgeb.org) or M.L. (marina.lusic@med.uni-heidelberg.de).

METHODS

Virus production. For the production of viral stocks, we used a plasmid obtained from the Env- molecular clone pNL4-3/E$^-$R$^-$, a gift from N. Landau. This viral clone harbours a frameshift mutation introduced near the 5′ end of the *env* gene[31] and performs a single-round infection once pseudotyped with vesicular stomatitis virus-G (VSV-G); this renders the virus incapable of spreading (and causing massive cell death).

We produced viral clone HIV-1_{BRU} as previously described[16]. The integrase (IN) defective packaging plasmid pcHelp/IN-, a gift from A. Cara, contains a D116N mutation in the IN genome, preventing the function of the IN protein[18]. The D64E mutant plasmid, which is similarly integration-defective, was obtained from the National Institutes of Health (NIH) AIDS Research and Reference Reagent Program[17].

Lentiviral vector pLV-THM[32] was obtained from Addgene, whereas pCCL18–green fluorescent protein (GFP) was modified from ref. 33 to become promoterless. The gammaretroviral MoMLV vector was a gift from G. Towers.

Infectious viral stocks were generated by transfecting viral DNA in HEK 293T cells and collecting supernatants after 48 h. Viral production was quantified by measuring viral p24 in the supernatants using the Innotest HIV antigen mAB kit (Innogenetics).

Primary cell isolation, culture and infection. Primary human CD4$^+$ T cells were isolated by Ficoll gradient separation, followed by purification with CD4 MicroBeads (Miltenyi Biotec). Cells were activated with a cocktail of beads containing 4.5×10^5 beads coated with αCD3 and αCD28 antibodies (Dynabeads Human T-Activator CD3/CD28 Dynal/Invitrogen), and plated in complete medium with interleukin-2 (IL-2; 30 U ml^{-1}, Sigma-Aldrich) for 4 days at 37 °C.

Activated CD4$^+$ T cells (1×10^6) were infected with 0.5–0.75 μg ml^{-1} of viral p24 for 4–5 h at 37 °C. After infection, cells were kept in culture at 1×10^6 cells per millilitre in complete RPMI 1640 medium supplemented with IL-2 and CD3/CD28 beads.

For raltegravir treatment, 10 μM raltegravir (obtained from the NIH AIDS Research and Reference Reagent Program) was added together with the virus during the infection, and it was later supplemented in the medium.

For generating the primary model of latency, naive CD4$^+$ T cells were isolated, cultured and infected as described in ref. 14.

Primary human CD34$^+$ cells were isolated, cultured and infected as described in ref. 34. Primary human macrophages were isolated and cultured as described in ref. 35.

Patients infected with HIV and having a CD4 count less than 3×10^5 ml^{-1} were enrolled before starting highly active antiretroviral therapy, following informed consent. Peripheral blood mononuclear cells obtained from healthy or infected blood donors according to a study protocol approved by the Ethical Committee of the Azienda Ospedaliero-Universitaria 'Ospedali Riuniti di Trieste', Italy, were isolated as described previously[16].

Cell culture and transfection. The Jurkat lymphoblastoid cell line, Jurkat J-Lat 15.4 clone and U937 monocytic cell line were kept in culture in complete RPMI 1640 medium with the addition of 10% fetal bovine serum (FBS). Cells were tested for mycoplasmal cell culture contaminants by using a MycoAlert kit from Lonza.

Transfection of Jurkat with p-eGFP–Nup153 expression plasmid[36] obtained from Euroscarf was done using Fugene HD (Promega), according the manufacturer's instructions. For the RNAi experiments, 3×10^6 cells were transfected with siRNA smart pools targeting LEDGF/p75 (PSIP1) (Dharmacon M-015209) or Nup153 (Dharmacon M-005283) proteins or with a non-targeting (NT) siRNA as a negative control (Dharmacon, Thermo Scientific). Transfection was performed with the Amaxa Nucleofection Device II (Amaxa), using an Amaxa nucleofection Kit V according to the manufacturer's instructions.

For western blot analysis, cells were harvested and homogenized in lysis buffer (20 mM Tris-HCl, pH 7.4, 1 mM EDTA, 150 mM NaCl, 0.5% Nonidet P-40, 0.1% SDS, 0.5% sodium deoxycholate) supplemented with protease inhibitors (Roche) for 10 min at 4 °C and sonicated (Bioruptor) for 5 min. Equal amounts of total cellular proteins (30 μg), as measured with Bradford reagent (Biorad), were resolved by 8% SDS–polyacrylamide gel electrophoresis (SDS–PAGE), transferred onto polyvinylidene difluoride membranes (GE Healthcare) and then probed with primary antibody (anti-Nup153 (Santa Cruz, sc-101544), anti-LEDGF/p75 (BD Biosciences, 611714), anti-GFP (LifeTechnologies, A6455)), followed by secondary antibody conjugated with horseradish peroxidase. The immunocomplexes were visualized with enhanced chemiluminescence kits (GE Healthcare).

Once protein silencing was assessed, cells were infected with 0.5–0.75 μg ml^{-1} of p24 of viral clone $NL_{4\text{-}3/E\text{-}R\text{-}}$ as described. At 24 and 48 h after infection, samples were collected for further analysis.

Integration assay (Alu PCR). Infected cells were tested for integration of HIV-1 by isolating genomic DNA from 1×10^6 cells with a DNeasy Tissue Kit (Qiagen). Genomic DNA (100 ng) was subjected to quantitative Alu-LTR PCR for integrated provirus or for 2-LTR circles as previously described[16].

Quantitative reverse transcription PCR. For the quantification of HIV transcript levels, RNA was purified from the cells with a Nucleospin RNA II purification kit (Macherey-Nagel). The messenger RNA (mRNA) levels were quantified by TaqMan quantitative reverse transcription PCR (qRT–PCR) using HIV-1 or interleukin-2 (IL-2) primers and probe[14], and housekeeping gene 18S and GAPDH as controls.

Luciferase activity assay. Cells were harvested 48 h after infection, and luciferase activity was measured using the Luciferase Assay Kit (Promega). Viral expression was expressed after normalization over micrograms of total cell extracts.

Cell preparation for three-dimensional immuno-DNA FISH. Three-dimensional FISH combined with immunostaining was performed according to protocols in ref. 37. Culture or primary cells were resuspended at 3×10^6 cells per millilitre in 5% FBS in PBS and cell suspension was allowed to attach to the glass cover slips, previously coated with poly-L-lysine. Cells were fixed in 4% paraformaldehyde in 0.3× hypotonic PBS for 10 min, permeabilized with PBS/0.5% Triton X-100 for 10 min and left in PBS/20% glycerol for 1 h. Cells were then blocked in PBS/5% fetal horse serum (FHS) for 45 min, and primary antibody (anti-NPC mAb414, Covance; anti-LaminB, Abcam ab16048; anti-GFP, LifeTechnologies A6455) was added for an overnight incubation at +4 °C in a humid chamber. The subsequent day, cells were washed five times in PBS-T (PBS with 0.05% Tween) and the secondary antibody (Jackson Laboratories) was used for 45 min at 22 °C (1/1,000 dilution). After five washings in PBS-T, cells were additionally crosslinked with EGS (ethylene glycol-bis(succinic acid *N*-hydroxysuccinimide ester) (Sigma E-3257) for 10 min, washed and permeabilized again in PBS-T/0.5% Triton X-100. After washing in PBS/0.05% Triton X-100, cells were rinsed and incubated in 0.1 N HCl (freshly prepared) for 10 min. Cells were left in PBS/20% glycerol for at least 45 min, and then subjected to five cycles of freeze and thaw in liquid nitrogen and PBS/20% glycerol. Additional washings in PBS/0.05% Triton X-100 preceded an overnight incubation in 50% formamide/2× SSC (hybridization buffer). The subsequent day, cells were treated with RNase A (100 μg ml^{-1} in 2× SSC) in a humid chamber at 37 °C for 1 h, were rinsed again in 50% formamide/2× SSC for at least 1 h (or overnight) and were then subjected to hybridization with the appropriate probe.

Probes for hybridization in three-dimensional immuno-DNA FISH. For visualization of the loci of interest, specific BAC clones (selected from CHORI (Children's Hospital Oakland Research Institute in Oakland, California) sites and purchased from Invitrogen) were isolated according to the manufacturer's instructions. The listing of the BACs with their identities and the genes they contain is provided in the Supplementary Information. BAC DNA (2 μg) was labelled with digoxigenin by Dig-Nick Translation (Roche) at 15 °C.

For the visualization of HIV-1, lentiviral or gamma-retroviral vector DNA integrated inside Jurkat or primary CD4$^+$ T cells, 2 μg of the respective plasmids was labelled by nick translation in the presence of 16-dUTP Biotin nucleotides at 15 °C for 3 h.

In both cases, probes were checked on agarose gel and then cleared by using an Illustra Microspin G-25 column (GE Healthcare) and precipitated in the presence of Cot-1 DNA (Roche) and DNA from herring sperm (Sigma). Finally, after ethanol precipitation, the probes were resuspended in 10 μl formamide, incubated at 37 °C for 15–20 min and 10 μl of 20% dextran in 4× SSC was added to a final volume of 20 μl.

Hybridization set up and development. The probe (1–10 μl) was loaded onto glass cover slips with the cells, followed by sealing with rubber cement, and heat-denatured on a heat block at 75 °C for 4 min. Hybridization was performed for 48 h at 37 °C in a humid chamber. Three washings in 2× SSC (10 min each) were followed with three washings in 0.5× SSC at 56 °C.

FISH development for Dig-labelled BACs was performed by using FITC-labelled anti-digoxigenin antibody (Roche), whereas biotin-labelled HIV-1 probes were detected by a TSA Plus system from Perkin Elmer, allowing signal amplification, by using an anti-biotin antibody (SA-HRP) and a secondary antibody with a fluorescent dye (usually FITC for HIV).

Microscopy. Three-dimensional-stacks of slides with fixed cells were captured on a Zeiss LSM 510 META confocal microscope (Carl Zeiss Microimaging) with a ×63 numerical aperture 1.4 Plan-Apochromat oil objective. The pinhole of the microscope was adjusted to obtain an optical slice of less than 1.0 μm for any wavelength acquired.

Distances observed between the FISH signals and the nuclear envelope were measured using LSM 510 Image Examiner Software (Zeiss) and Volocity (Perkin Elmer); measurements were normalized over nuclear radius (defined as half of the middle of the mAb414-TRITC ring), and then binned into three classes of equal surface area[11].

Measurements were acquired for the alleles of the following genes in activated CD4$^+$ T cells: *NPLOC4* ($n = 90$), *FKBP5* ($n = 128$), *NFATC3* ($n = 68$), *HEATR7A* ($n = 78$), *RPTOR* ($n = 94$), *SPTAN* ($n = 72$), *SMG1* ($n = 80$), *GRB2* ($n = 132$), *KDM2A* ($n = 102$), *DNMT1* ($n = 136$), hotter zone 1 ($n = 118$), hotter zone 2 ($n = 100$), hotter zone 3 ($n = 62$), hotter zone 4 ($n = 154$), *GAPDH* ($n = 52$), *ACTN1* ($n = 126$),

CD4 ($n = 100$), *CD28* ($n = 52$), *HEATR6* ($n = 44$), *KDM2B* ($n = 106$), *PACS2* ($n = 42$), *IKZF3* ($n = 52$), *TAP2* ($n = 54$), *CNTN4* ($n = 56$), *GPC5* ($n = 30$), *PTPRD* ($n = 30$). Measurements were acquired for the alleles of the following genes in $CD34^+$ cells: *IKZF3* ($n = 54$) and *TAP2* ($n = 60$). Measurements were acquired for the proviruses in primary macrophages ($n = 18$) and the U937 cell line ($n = 30$). Measurements were acquired for the proviruses in primary cells upon several conditions: 4 days after infection of activated $CD4^+$ T cells ($n = 160$ HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$ and $n = 42$ HIV-1_{BRU} measured in three independent experiments); HIV-1 in $CD4^+$ T cells from infected patients ($n = 28$ and $n = 27$); 4 days after infection with mutant viruses IN(D64E) ($n = 30$) or IN(D116N) ($n = 66$), or cells infected with HIV-$1_{NL4.3}$ upon raltegravir treatment ($n = 159$); latent $CD4^+$ cells (2 weeks after infection) with or without CD3/CD28 stimulation ($n = 40$ and $n = 33$, respectively). Measurements were acquired for proviruses in Jurkat or J-Lat cell lines 4 days after infection of Jurkat with HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$, in different conditions: no transfection (control, $n = 116$), transfections with non-targeting siRNA (siNT2/NT5, $n = 163$), LEDGF/p75 siRNA (siLEDGF, $n = 164$), Nup153 (siNup153, $n = 129$), Nup153 siRNA + enhanced GFP (eGFP)–Nup153 ($n = 52$). The corresponding graphs show the average results from three independent experiments. Proviral DNA was analysed in the J-Lat clone 15.4 with and without TPA ($n = 74$ and $n = 150$, respectively); Jurkat + Lentiviral promoter ($n = 19$); Jurkat + transcription-less lentiviral vector ($n = 51$); Jurkat + gammaretroviral vector ($n = 88$).

ChIP. $CD4^+$ T cells (40×10^6) were washed twice in PBS before crosslinking with 1% final formaldehyde for 10 min at room temperature, followed by termination of the reaction with 125 mM glycine on ice. The cell pellet was washed twice with PBS and was lysed in 0.5% NP-40 buffer (10 mM Tris-Cl pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1 mM PMSF and protease inhibitors). The nuclei obtained were washed once in the same buffer without NP-40. Lysis of the nuclei was performed using the same buffer containing 4% of NP-40 at 37 °C for 15 min, upon which micrococcal nuclease was added (120 units of the enzyme), and the reaction was stopped with 3 mM EGTA. DNA was additionally sheared by sonication to an average size of DNA fragments below 500 base pairs. Extracts were pre-cleared by two rounds of incubation with immunoglobulin-γ and agarose beads, followed by centrifugation at 1,200*g* for 5–10 min. The lysate (400 µl) was then incubated with 2–4 µg of the indicated antibody overnight at 4 °C, followed by incubation for 4 h with MagnaChIP Protein A/G Magnetic Beads (Millipore). Beads were then washed thoroughly with RIPA150, with LiCl-containing buffer and with TE, RNase treated for at least 30 min at 37 °C, and treated with proteinase K for at least 2 h at 56 °C. De-crosslinking of protein–DNA complexes was performed by an overnight incubation at 65 °C. DNA was then extracted by phenol–chloroform extraction followed by ethanol precipitation and was quantified by real-time PCR. The following antibodies were used in the ChIP experiments: mAb414 (Covance, MMS-120R), anti-Pol2 (Santa Cruz, sc-9001X), anti-USF1 (Santa Cruz, sc-229X), anti-NF-κB p65 subunit (Santa Cruz, sc-109X), anti-Nup153 [SA1] (Abcam, ab-96462) and anti-Nup153 (QE5) (Abcam, ab-24700), anti-Nup98 (Cell Signaling, 2598), anti-Nup62 (BD Biosciences, 610497), anti-Tpr (Abcam ab-58344), anti-Mcm2 (Abcam), mouse IgG (Santa Cruz, sc-2025). The graphs of ChIP figures show the mean and s.e.m. from at least three independent experiments.

Bioinformatics and statistical analysis. No statistical methods were used to predetermine sample size.

Five lists of HIV-1 integration sites were collected from published work[2,3,38–40] and an unpublished list of integration sites in $CD4^+$ T cells provided by A.R. and F.M. (Extended Data Table 1). HIV-1 RIGs ($n = 156$) were genes found in more than one list (Supplementary Information and Extended Data Table 1); their genomic position was plotted onto the chromosome map using the Idiographica webtool (http://www.ncrna.org/idiographica (ref. 9); Extended Data Fig. 2). Genomic coordinates of eight selected hotter zones, into which HIV-1 integration density was found higher than expected, were downloaded from the Bushman Lab website (http://www.bushmanlab.org/tutorials/ucsc)[1] and reported in the Supplementary Information.

The calculation of the probability of finding 156 genes present in more than one list by chance was performed by computer simulation. A program was written to randomly draw, from 25,000 genes, 265, 329, 294, 32, 158 and 58 genes, and to count the genes drawn more than once. The simulation was repeated 1×10^9 times. The distribution obtained is shown in Extended Data Fig. 1. Calculations were performed using the Matlab 2011R software (http://www.mathworks.com/).

Expression of HIV RIGs and control genes was derived from published transcriptomic data in $CD4^+$ T cells, using biogps.org (as in refs 12, 13, 22), and was compared with a random sets of genes. Using the non-parametric Mann–Whitney-Wilcoxon test, it was concluded that that HIV RIGs and control genes are more transcribed than a random sets of genes ($P < 0.005$).

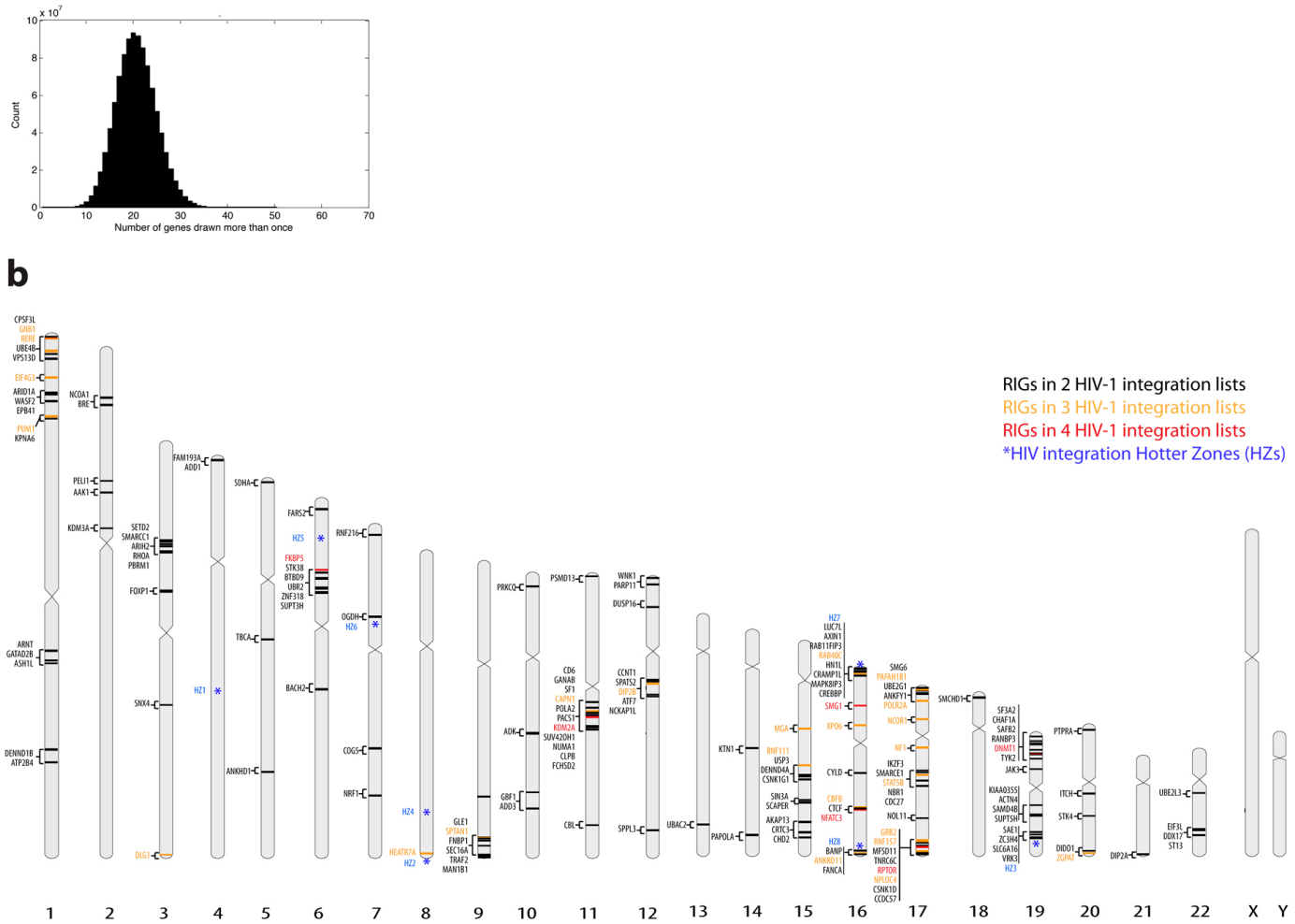
LAD coordinates were obtained from ref. 5, whereas genes inside LADs were derived from BioMart Ensembl and named LAD genes. Then, the *P* value of the common genes in the LAD genes, HIV RIGs, the six integration lists and cold genes were calculated by pairwise comparison of each combination, followed by hypergeometric test (Fig. 3g). The following *P* values were obtained: ref. 3, $P = 1.27 \times 10^{-8}$; ref. 2, $P = 0.008$; ref. 38, $P = 0.001$; ref. 40, $P = 0.36$; ref. 39, $P = 0.0007$; F.M. *et al.* (unpublished observations), $P = 1.32 \times 10^{-15}$; ref. 1 hotter zones, $P = 0.0003$; HIV RIGs, $P = 2.09 \times 10^{-10}$. Eighty per cent of genes that are never targeted by HIV-1 (cold genes) are significantly enriched inside LADs ($P = 3.25 \times 10^{-19}$).

The profile of aligned LAD border regions (Fig. 3i) was performed as described in ref. 5. A χ^2 test was applied to compare the distribution across the LAD border of HIV RIGs with the one of 3,000 random genes that were generated without replacement using the RSA-tool (http://floresta.eead.csic.es/rsat/random-genes_form.cgi)[41].

ChIP-seq profile analyses (Fig. 3b–f and Extended Data Fig. 6) were performed as in refs 22, 42. The 1,000 most expressed and 1,000 least expressed genes were obtained as in ref. 22. and named active and silent genes, respectively. The TSS coordinates of these genes were obtained using the University of California, Santa Cruz (UCSC) Table Browser.

Comparison between groups for expression data was performed using a non-parametric Mann–Whitney–Wilcoxon rank sum test; comparison of gene distributions was by a χ^2 test, with the exception of data shown in Fig. 3g, which were analysed by a hypergeometric test. For the FISH, ChIP and real-time PCR results, the reported values are means and s.e.m., calculated from at least three independent samples. For statistical comparison of three or more groups, one-way analysis of variance followed by Tukey's post-hoc test was used. A value of $P < 0.05$ was considered significant.

31. Connor, R. I., Chen, B. K., Choe, S. & Landau, N. R. Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. *Virology* **206,** 935–944 (1995).
32. Wiznerowicz, M. & Trono, D. Conditional suppression of cellular genes: lentivirus vector-mediated drug-inducible RNA interference. *J. Virol.* **77,** 8957–8961 (2003).
33. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *J. Virol.* **72,** 8463–8471 (1998).
34. Cattoglio, C. *et al.* Hot spots of retroviral integration in human $CD34^+$ hematopoietic cells. *Blood* **110,** 1770–1778 (2007).
35. Repnik, U., Knezevic, M. & Jeras, M. Simple and cost-effective isolation of monocytes from buffy coats. *J. Immunol. Methods* **278,** 283–292 (2003).
36. Daigle, N. *et al.* Nuclear pore complexes form immobile networks and have a very low turnover in live mammalian cells. *J. Cell Biol.* **154,** 71–84 (2001).
37. Solovei, I. & Cremer, M. 3D-FISH on cultured cells combined with immunostaining. *Methods Mol. Biol.* **659,** 117–126 (2010).
38. Han, Y., Wind-Rotolo, M., Yang, H. C., Siliciano, J. D. & Siliciano, R. F. Experimental approaches to the study of HIV-1 latency. *Nature Rev. Microbiol.* **5,** 95–106 (2007).
39. Ikeda, T., Shibata, J., Yoshimura, K., Koito, A. & Matsushita, S. Recurrent HIV-1 integration at the BACH2 locus in resting $CD4^+$ T cell populations during effective highly active antiretroviral therapy. *J. Infect. Dis.* **195,** 716–725 (2007).
40. Liu, H. *et al.* Integration of human immunodeficiency virus type 1 in untreated infection occurs preferentially within genes. *J. Virol.* **80,** 7765–7768 (2006).
41. van Helden, J. Regulatory sequence analysis tools. *Nucleic Acids Res.* **31,** 3593–3596 (2003).
42. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138,** 1019–1031 (2009).



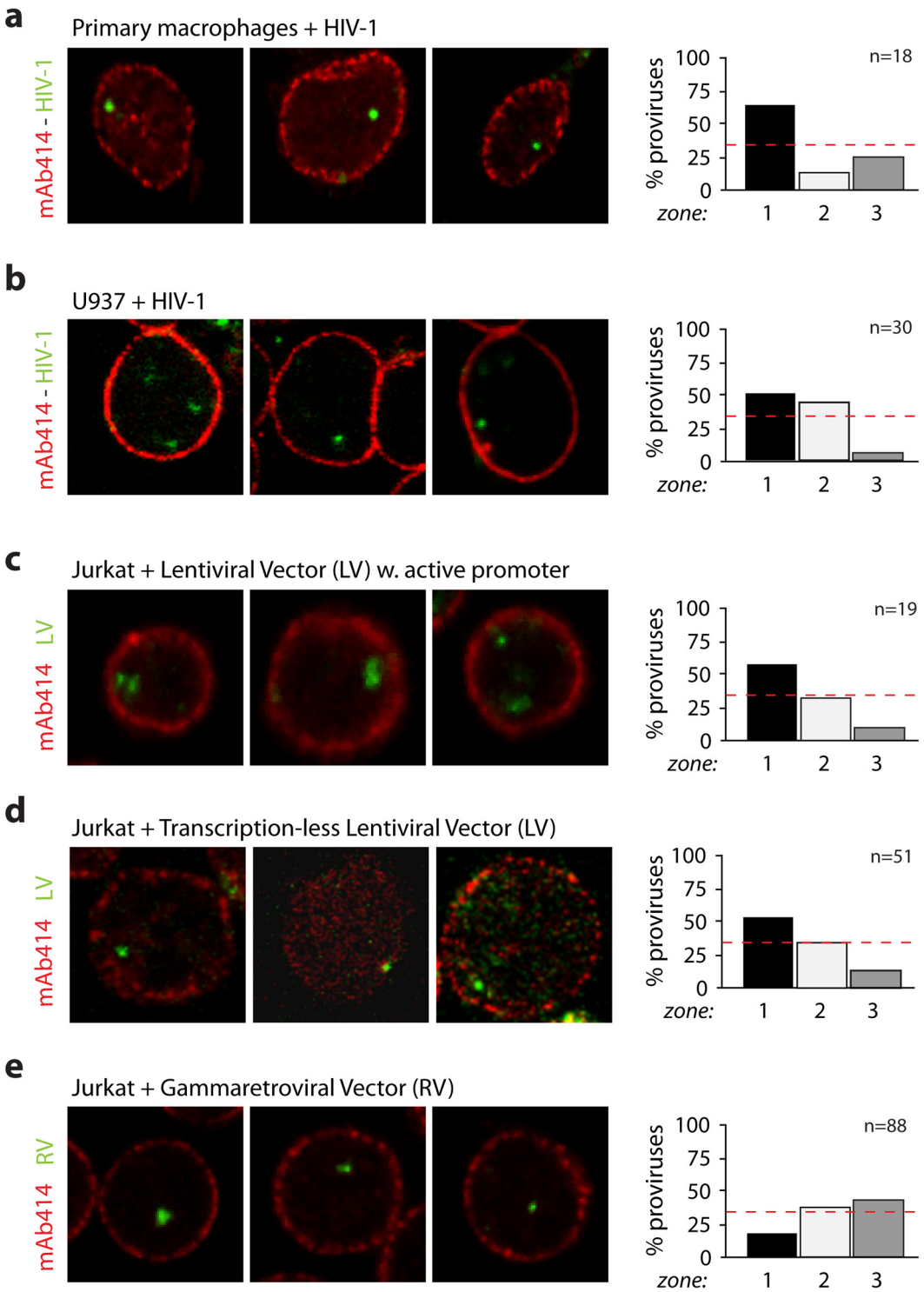
Extended Data Figure 1 | HIV-1 RIGs. **a**, Probability of recurrence of a random set of genes in different lists of HIV-1 integration sites. The histogram shows the distribution of the number of genes present at least twice in the six HIV-1 integration site lists considered; 1×10^9 independent drawings were evaluated. The distribution peaks around 20 genes, with a maximum observed of 50. The number of RIGs detected experimentally in at least two lists was instead 156 ($P < 1 \times 10^{-9}$). **b**, Human chromosome map showing the localization of 156 HIV RIGs. Genes found in four, three and two HIV-1 integration lists are highlighted in red, orange and black, respectively. Hotter genomic regions, favoured for HIV-1 integration as described in ref. 1, are highlighted in blue and indicated by a star.

LOCUS	within	1 Mb	5 Mb	10 Mb
6p21.31	RIGs	1	2	5
	other integration sites	1	10	16
	TOT.	2	12	21
9q34.11	RIGs	1	2	5
	other integration sites	5	7	12
	TOT.	6	9	17
11q13.2	RIGs	0	7	10
	other integration sites	0	13	17
	TOT.	0	20	27
16p12.3	RIGs	0	0	1
	other integration sites	0	6	11
	TOT.	0	6	12
16q22.1	RIGs	2	2	2
	other integration sites	3	6	7
	TOT.	5	8	9
17q25.1-3	RIGs	3	5	6
	other integration sites	11	14	16
	TOT.	14	19	22
19p13.2	RIGs	1	3	6
	other integration sites	5	13	29
	TOT.	6	16	35
8q24.3	RIGs	0	0	0
	other integration sites	0	0	0
	TOT.	0	0	0
	TOT.	**33**	**90**	**139**

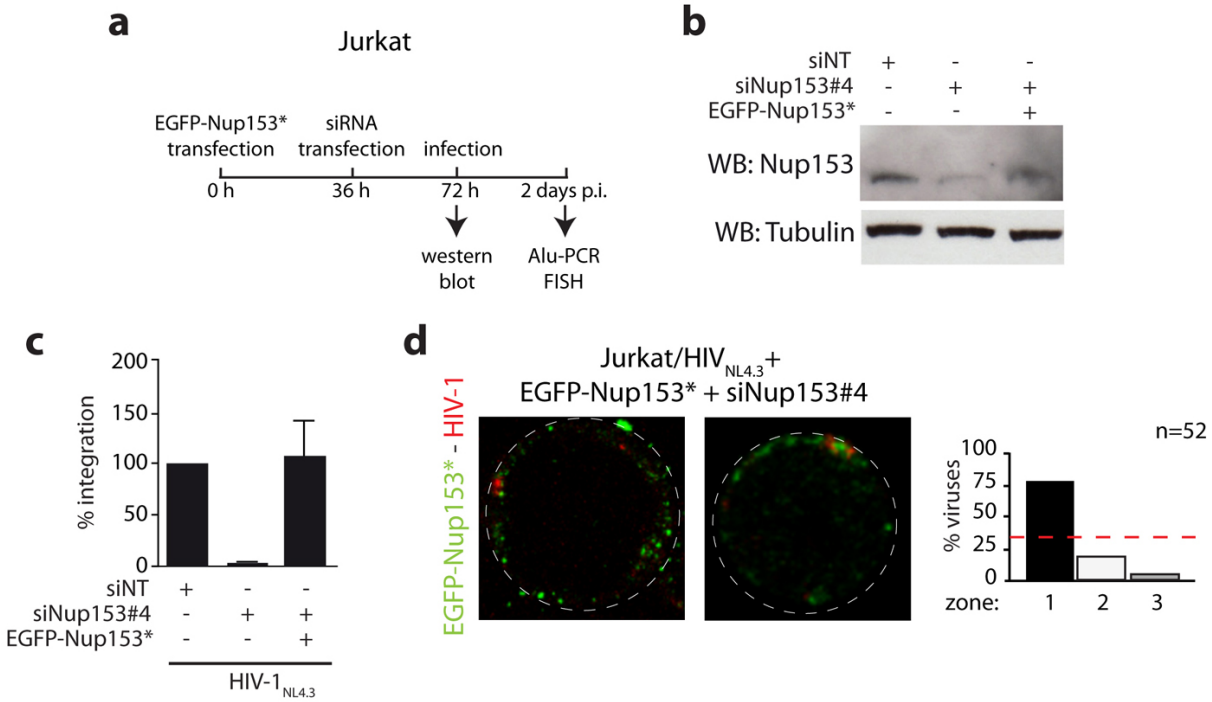
LOCUS	within	1 Mb	5 Mb	10 Mb
4q25	RIGs	0	0	0
	other integration sites	1	4	5
	TOT.	1	4	5
8q24.3	RIGs	1	1	1
	other integration sites	7	9	9
	TOT.	8	10	10
19q13.42	RIGs	0	0	1
	other integration sites	0	5	7
	TOT.	0	5	8
8q24.21	RIGs	0	0	0
	other integration sites	2	7	7
	TOT.	2	7	7
	TOT.	**11**	**26**	**30**

within	1 Mb	5 Mb	10 Mb
total number of RIGs/integrants in the proximity of the BAC probes	**44**	**116**	**169**

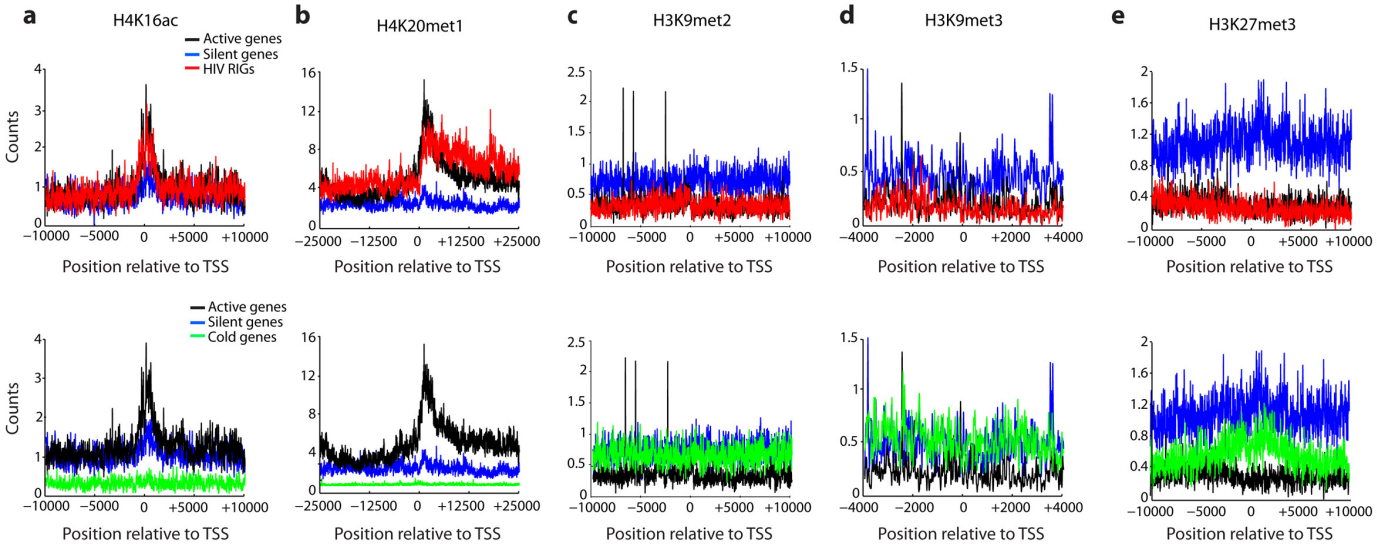
Extended Data Figure 2 | Distribution of RIGs or individual integration sites all over the loci analysed by FISH. The scheme describes the distribution of RIGs (bold) or simple integration sites (regular) around the locus analysed by FISH in Fig. 1; RIGs are in the left panel and hotter zones are in the right panel. As indicated on the side, the total number of RIGs/integrants was calculated within 1, 5 or 10 Mb from the locus analysed by FISH. In total, considering all the RIGs and hotter zones, there are 44 other RIGs/integrants within a window of 1 Mb, 116 within 5 Mb and 169 within 10 Mb around the analysed locus.



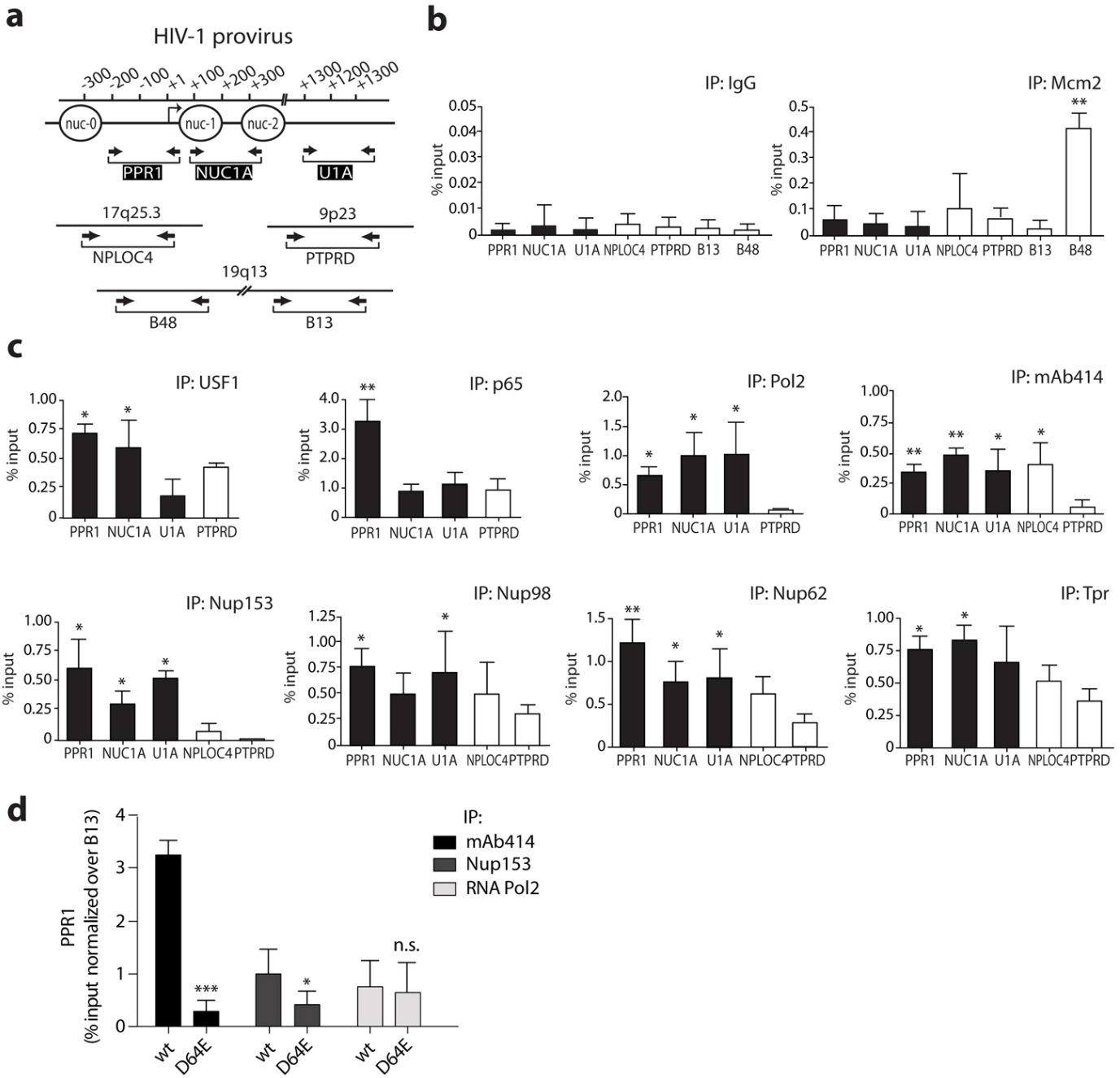
Extended Data Figure 3 | FISH analysis. **a**, Representative images of three-dimensional immuno-DNA FISH of HIV-1 DNA (green) in human primary HIV-1 macrophages stained for mAb414 (red), with relative distribution of FISH signals according to the three concentric zones. **b**, FISH of HIV-1 DNA (green) in the HIV-1 infected U937 monocytic cell line. **c**, Representative images of three-dimensional immuno-DNA FISH of lentiviral vector pLV-THM (green) in Jurkat cells. **d**, Representative images of three-dimensional immuno-DNA FISH of the promoter-less lentiviral vector pCCL–18GFP (green) in Jurkat cells. **e**, Representative images of three-dimensional immuno-DNA FISH of a gammaretroviral vector (green) in Jurkat cells. For all panels, the graphs are organized as described in the main text.



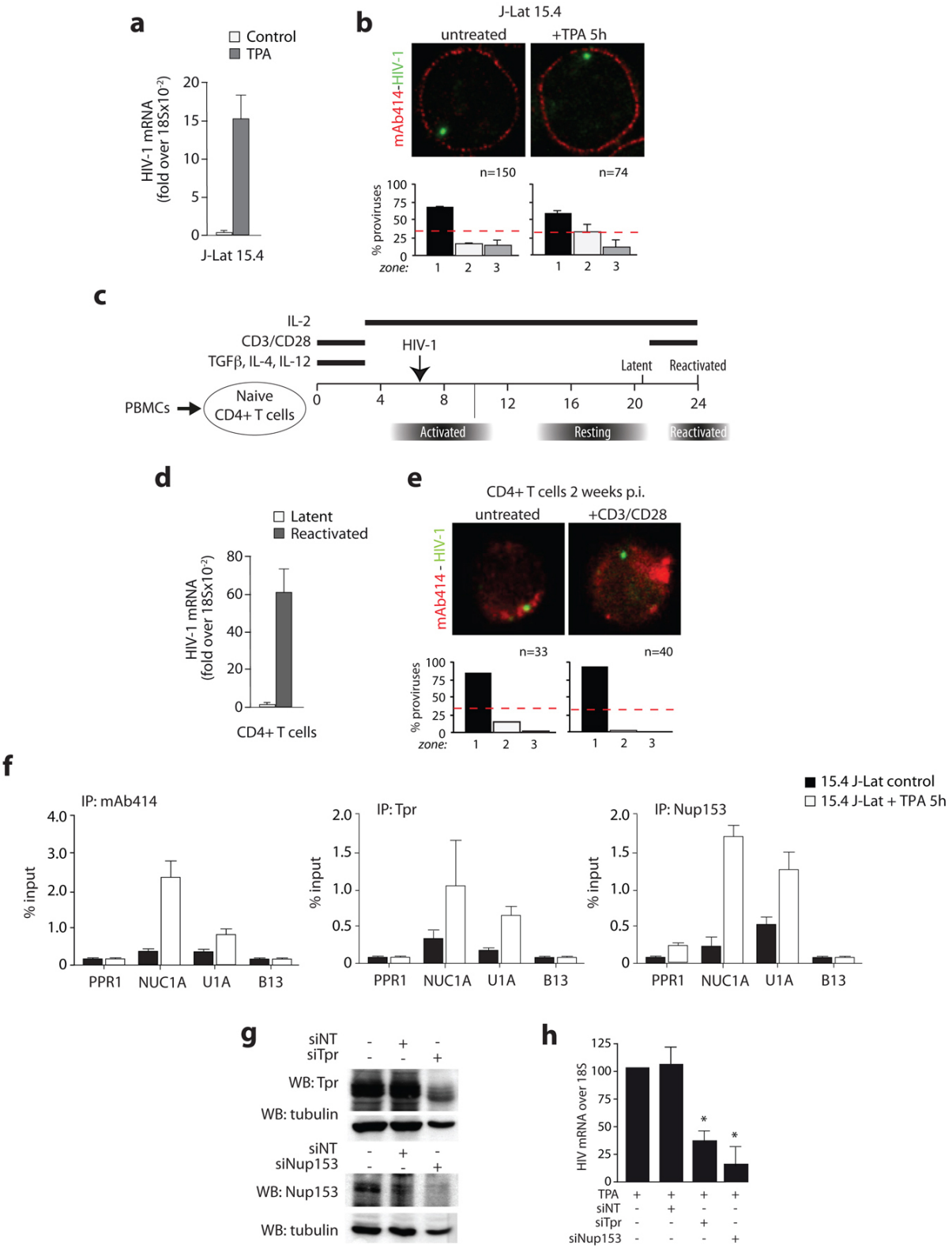
Extended Data Figure 4 | Reconstitution of Nup153 by transfection of an siRNA-resistant plasmid coding for eGFP–Nup153. **a**, Scheme of the experiment performed in Jurkat cells. eGFP–Nup153* contains the coding region for Nup153 tagged with eGFP, but is devoid of the 3′ untranslated region of the mRNA, which is the target of the anti-Nup153 siRNA4. **b**, Western blot showing Nup153 protein level at the moment of infection. siNT, not targeting siRNA. **c**, Real-time Alu-PCR in Jurkat cells 2 days after infection with HIV-$1_{NL4.3}$. Values are mean and s.e.m. of three experiments after normalization over Jurkat transfected with a control, non-targeting siRNA (siNT). **d**, Representative images of three-dimensional immuno-DNA FISH of HIV-1 DNA (red) in Jurkat cells transfected first with the eGFP–Nup153* expression plasmid and then with the siRNA4, targeting endogenous Nup153. The graph on the right side shows the distribution of HIV-1 FISH signals according to the three concentric zones in cells expressing eGFP.



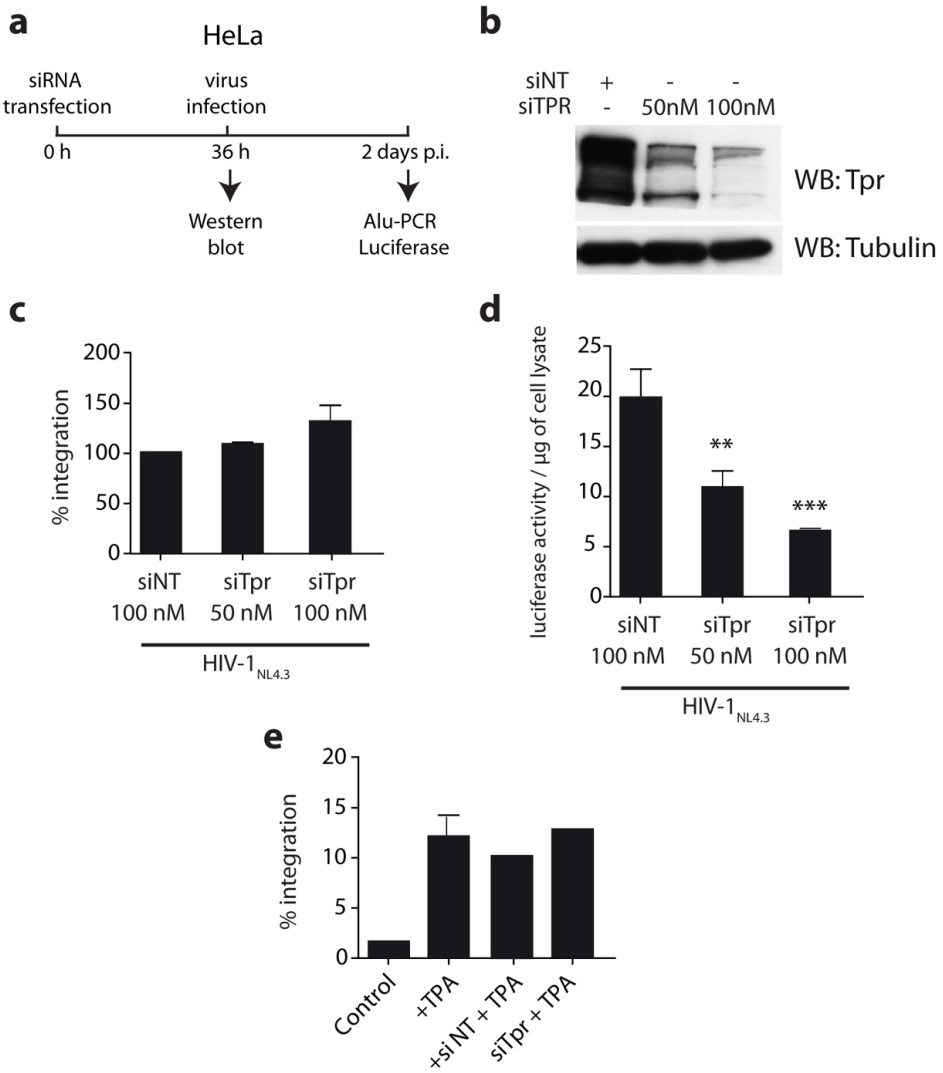
Extended Data Figure 5 | ChIP-seq profiles for HIV RIGs, cold genes and controls. **a**–**e**, Profiles of chromatin modifications around the TSS for HIV RIGs (red) and cold genes (green) compared with highly active (black) and silent (blue) genes in activated $CD4^+$ T cells. Each panel reports results for a specific modification, as indicated.



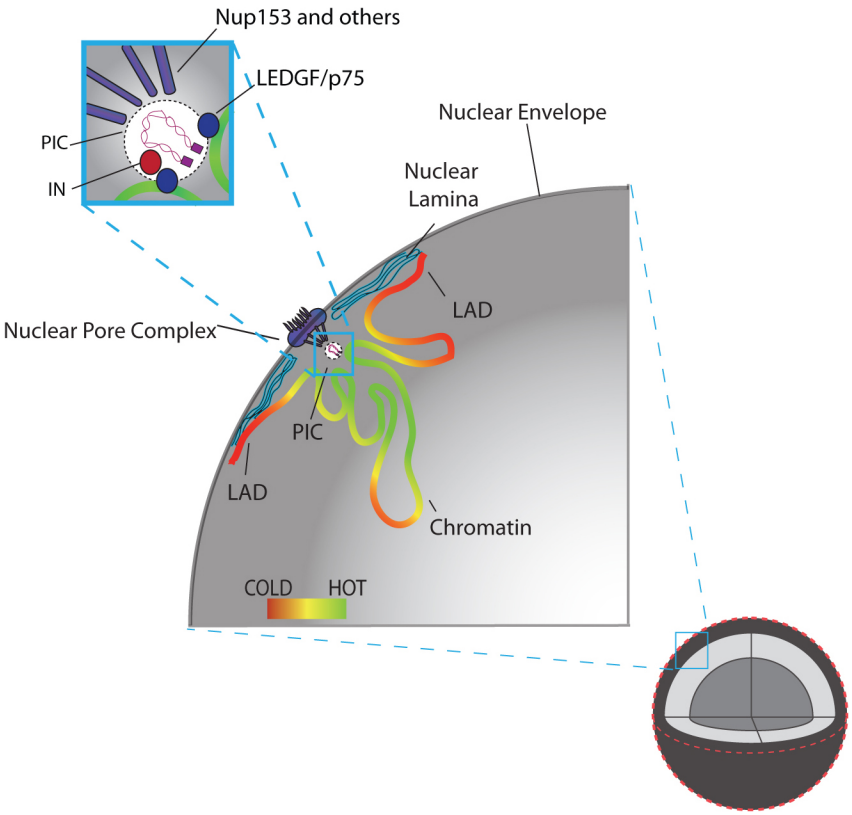
Extended Data Figure 6 | Association of HIV-1 provirus with nucleoporins. **a**, Positions of primers used for ChIP on the HIV-LTR (numbering is according to the TSS and nucleosomes are shown), the NPLOC4 RIG, the PTPRD cold gene, B48 and B13 genomic controls for DNA standardization. B48 maps within the human lamin B2 origin of DNA replication. **b**, Control ChIP data in $CD4^+$ T cells infected with HIV-$1_{NL4\text{-}3/E\text{-}R\text{-}}$, using total immunoglobulin-γ and an antibody against the unrelated Mcm2 cellular protein. For each analysed region, the amount of immunoprecipitated chromatin using the indicated antibodies was normalized according to the input amount of chromatin. Mean and s.e.m. from at least three independent experiments. $^{**}P < 0.01$. **c**, ChIP results in $CD4^+$ T cells, 4 days after HIV-1 infection, using the indicated antibodies. The amount of immunoprecipitated chromatin was normalized according to input. Mean and s.e.m. from at least three independent experiments. $^{**}P < 0.01$,$^{*}P < 0.05$. **d**, ChIP results in $CD4^+$ T cells, 4 days after infection with wild-type HIV-1 or the IN(D64E) mutant virus. For the PPR1 region, corresponding to the viral promoter, the amount of immunoprecipitated chromatin using the indicated antibodies (mAb414, Nup153 and Pol2) was calculated according to the input amount of chromatin, and then normalized over the B13 control genomic region. The graphs show the mean and s.e.m. from three independent experiments. $^{***}P < 0.001$; $^{*}P < 0.05$.



Extended Data Figure 7 | HIV-1 transcriptional activation is concomitant with, and requires, nucleoporins binding to the provirus. **a**, Quantitative reverse transcription PCR measurement of HIV-1 mRNA in mock- or TPA-treated J-Lat 15.4 cells. **b**, Three-dimensional immuno-DNA FISH of HIV-1 DNA (green) in J-Lat 15.4 cells stained for NPC (red) before and after TPA reactivation. **c**, Scheme of the experiment for the generation of a primary, cellular model of HIV-1 latency to study HIV-1 DNA localization in activated and resting primary $CD4^+$ T cells. **d**, Quantitative reverse transcription PCR measurement of HIV-1 mRNA levels in primary infected $CD4^+$ T cells before and after reactivation, normalized over the 18S housekeeping gene. Latent versus reactivated: $P < 0.001$. **e**, Three-dimensional immuno-DNA FISH of HIV-1 DNA (green) in latently infected $CD4^+$ T cells stained for the NPC (red) before and after reactivation, with relative distribution of HIV-1 FISH signals according to the three concentric zones considered in this work. **e**, ChIP in control and TPA-stimulated J-Lat 15.4 cells, with the indicated antibodies. Mean and s.e.m. from at least three independent experiments. **g**, Immunoblot for Tpr (upper panel) and Nup153 (lower panel), 36 h after transfection of the indicated siRNAs (NT, non-targeting control). **h**, Levels of HIV-1 RNA in siRNA-treated J-Lat 15.4 cells after TPA activation. Mean and s.e.m. from three independent experiments. *$P < 0.05$.



Extended Data Figure 8 | Silencing of Tpr in HeLa cells and 15.4 J-Lat clones. **a**, Scheme of the experiment to study HIV-1 integration in infected HeLa cells after Tpr silencing. **b**, Western blot showing Tpr protein level at the moment of infection, after treatment of HeLa cells with a non-targeting siRNA (siNT) or an siRNA targeting Tpr at two different doses. Values are mean and s.e.m. of three experiments after normalization over HeLa cells transfected with a control non-targeting siRNA. **c**, Real-time Alu PCR in HeLa cells infected with HIV-$1_{NL4.3}$ and previously transfected with a non-targeting siRNA (siNT) or an siRNA targeting Tpr at two different doses. Values are mean and s.e.m. of three experiments after normalization over HeLa cells transfected with a control non-targeting siRNA. **d**, Luciferase activity assay in HeLa infected with HIV-$1_{NL4.3}$ and previously transfected with a non-targeting siRNA (siNT) or an siRNA targeting Tpr at two different doses. Values are mean and s.e.m. of three experiments. Statistical significance: $***P < 0.001$; $**P < 0.01$. **e**, Real-time PCR quantification of IL-2 mRNA levels in J-Lat 15.4 cells. The following conditions were tested: untreated cells, plus TPA (4 h), transfection with non-targeting siRNA or an siRNA targeting Tpr for 24 h, followed by treatment with TPA (4 h). Values are mean and s.e.m. of three experiments after normalization over GAPDH. Transcription of interleukin-2 (IL-2) was not significantly altered upon Tpr downregulation.



Extended Data Figure 9 | Model for HIV-1 integration site selection. After entry into the nucleus through the nuclear pore, the viral DNA integrates into the active chromatin closest to the NPC (green zones), avoiding both LADs and the inner part of the nucleus (red zones).

Extended Data Table 1 | List of HIV-1 integration sites considered in this work

List	Source	Nr. Published Sequences	Nr. Unique Intragenic Sites	Reference
Brady et al.	Primary, activated CD4+ T cells, in vitro infection	524	265	(Brady et al., 2009)
Mavilio et al.	Primary, activated CD4+ T cells, in vitro infection	638	329	Unpublished
Schroeder et al.	Sup T1, in vitro infection	642	294	(Schroder et al., 2002)
Liu et al.	PBMCs and tissues from HIV patients	42	32	(Liu et al., 2006)
Ikeda et al.	CD4+ T cells from HIV patients	463	158	(Ikeda et al., 2007)
Han et al.	CD4+ T cells from HIV patients	74	58	(Han et al., 2004)
TOTAL			**1136**	
N. Genes in 4 lists			6	
N. Genes in 3 lists			24	
N. Genes in 2 lists			126	
TOTAL N. RECURRENT GENES			**156**	

Out of the indicated numbers of sequences identified by the six considered studies, 1,136 were within individual genes; of these, 156 recurred in two or more studies.

LETTER

doi:10.1038/nature14443

The *Xist* lncRNA interacts directly with SHARP to silence transcription through HDAC3

Colleen A. McHugh[1]*, Chun-Kan Chen[1]*, Amy Chow[1], Christine F. Surka[1], Christina Tran[1], Patrick McDonel[2], Amy Pandya-Jones[3,4], Mario Blanco[1], Christina Burghard[1], Annie Moradian[5], Michael J. Sweredoski[5], Alexander A. Shishkin[1], Julia Su[1], Eric S. Lander[2], Sonja Hess[5], Kathrin Plath[3,4] & Mitchell Guttman[1]

Many long non-coding RNAs (lncRNAs) affect gene expression[1], but the mechanisms by which they act are still largely unknown[2]. One of the best-studied lncRNAs is *Xist*, which is required for transcriptional silencing of one X chromosome during development in female mammals[3,4]. Despite extensive efforts to define the mechanism of *Xist*-mediated transcriptional silencing, we still do not know any proteins required for this role[3]. The main challenge is that there are currently no methods to comprehensively define the proteins that directly interact with a lncRNA in the cell[5]. Here we develop a method to purify a lncRNA from cells and identify proteins interacting with it directly using quantitative mass spectrometry. We identify ten proteins that specifically associate with *Xist*, three of these proteins—SHARP, SAF-A and LBR—are required for *Xist*-mediated transcriptional silencing. We show that SHARP, which interacts with the SMRT co-repressor[6] that activates HDAC3[7], is not only essential for silencing, but is also required for the exclusion of RNA polymerase II (Pol II) from the inactive X. Both SMRT and HDAC3 are also required for silencing and Pol II exclusion. In addition to silencing transcription, SHARP and HDAC3 are required for *Xist*-mediated recruitment of the polycomb repressive complex 2 (PRC2) across the X chromosome. Our results suggest that *Xist* silences transcription by directly interacting with SHARP, recruiting SMRT, activating HDAC3, and deacetylating histones to exclude Pol II across the X chromosome.

Over the last two decades, numerous attempts have been made to define the protein complexes that interact with *Xist* and that are required for its various roles in X-chromosome inactivation (XCI)[3]. Most studies have used prior knowledge of the molecular events that occur on the X chromosome to define potential *Xist*-interacting proteins[8,9]. Although individual proteins that associate with *Xist* have been identified[8,10], we still do not know any of the proteins required for *Xist*-mediated transcriptional silencing because perturbations of these proteins, including components of the PRC2 complex, have no effect on *Xist*-mediated transcriptional silencing[11,12]. Current methods for identifying lncRNA-interacting proteins either require selection of specific candidate interacting proteins or fail to distinguish between direct RNA interactions that occur in the cell from those that merely associate in solution (reviewed in ref. 5).

To develop a method for identifying the proteins that directly interact with a specific lncRNA *in vivo*, we adapted our RNA antisense purification (RAP) method[13] to purify a lncRNA complex and identify the interacting proteins by quantitative mass spectrometry (RAP-MS) (Methods, Fig. 1a). Briefly, RAP-MS uses ultraviolet (UV) crosslinking to create covalent bonds between directly interacting RNA and protein and purifies lncRNAs in denaturing conditions to disrupt non-covalent interactions (Methods). This UV-crosslinking and denaturing approach, which is used by methods such as crosslinking and immunoprecipitation (CLIP), is known to identify only direct RNA–protein interactions and to separate interactions that are crosslinked in the cell from those that merely associate in solution[5,14].

Adapting this UV-crosslinking and denaturing approach to enable purification of a specific lncRNA is challenging for several reasons. (1) To purify lncRNA complexes in denaturing conditions, we need an RNA capture method that can withstand harsh denaturing conditions. (2) To detect the proteins associated with a given lncRNA, we need to achieve high purification yields of a lncRNA complex because, unlike nucleic acids, we cannot amplify proteins before detection. (3) Because any individual RNA is likely to be present at a very low percentage of the total cellular RNA, we need to achieve high levels of enrichment to identify specific interacting proteins. (4) Because the number of background proteins will be high, even after enrichment, we need accurate and sensitive methods for protein quantification to detect specific lncRNA-interacting proteins.

The RAP-MS method addresses these challenges because (1) RAP uses long biotinylated antisense probes, which form very stable RNA–DNA hybrids, and therefore can be used to purify lncRNA complexes in denaturing and reducing conditions (that is, 4 M urea at 67 °C, see Methods). (2) We optimized the RAP method to achieve high yields of endogenous RNA complexes. In our original protocol[13], we achieved <2% yield of the endogenous RNA complex; by optimizing hybridization, washing, and elution conditions (Methods), we were able to reproducibly achieve ~70% yield (Extended Data Fig. 1a, Methods). (3) Using our optimized conditions, we increased the enrichment levels for the target lncRNA complex (~5,000-fold, Extended Data Fig. 1b) relative to our already high levels of enrichment achieved previously (~100-fold)[13]. (4) To achieve sensitive quantification and to distinguish between specific proteins and background proteins, we used stable isotope labelling by amino acids in culture (SILAC) to label proteins (Methods, Extended Data Fig. 1c), which enables quantitative comparisons of purified proteins by mass spectrometry[15].

We validated the RAP-MS approach by defining the proteins that interact with two well-characterized non-coding RNAs: *U1* (a core component of the spliceosome) and *18S* (a component of the small ribosomal subunit). In the *U1* purifications, we identified 9 enriched proteins, all of which are known to interact with *U1* (Supplementary Note 1). In the *18S* purification, we identified 105 enriched proteins; 98 of these (93%) were previously characterized as ribosomal proteins, ribosomal processing and assembly factors, translational regulators, or other known ribosome interactors (Extended Data Fig. 2). In particular, we identified 21 of the 31 known small ribosomal subunit proteins. The few missing proteins appear to fall predominately into two categories: proteins that make few direct contacts with the RNA and small proteins that contain few peptides that could be detected by mass spectrometry.

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA. [3]Department of Biological Chemistry, Jonsson Comprehensive Cancer Center, Molecular Biology Institute, University of California Los Angeles, Los Angeles, California 90095, USA. [4]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA. [5]Proteome Exploration Laboratory, Beckman Institute, California Institute of Technology, Pasadena, California 91125, USA.
*These authors contributed equally to this work.

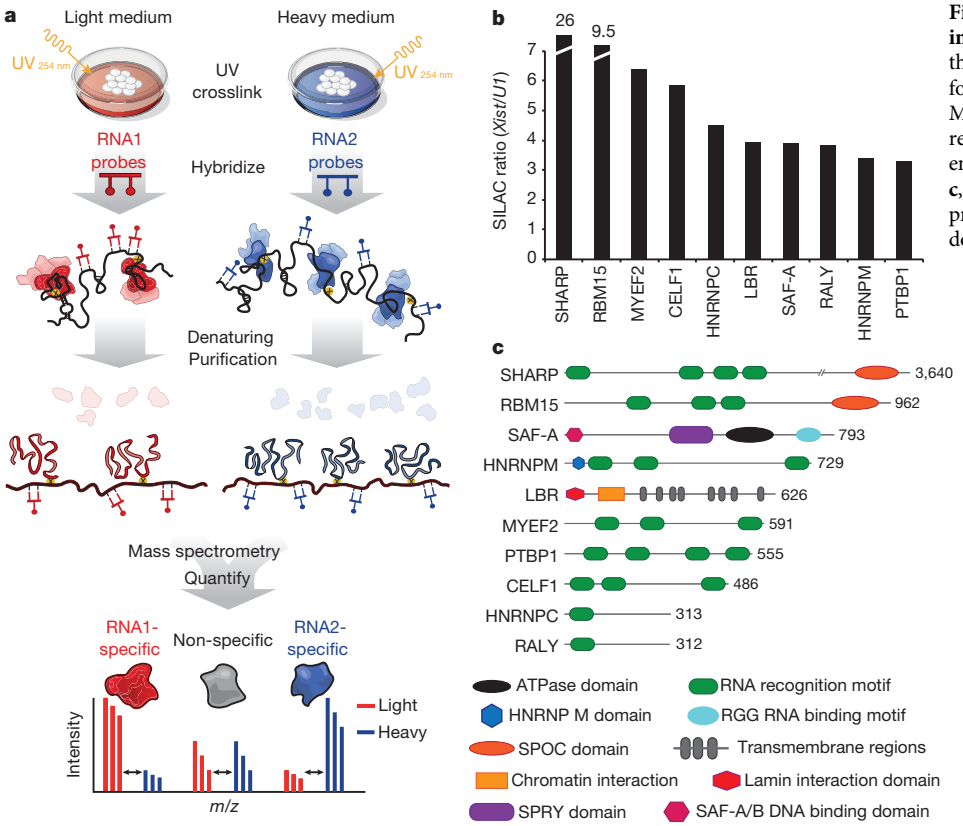


Figure 1 | RAP-MS identifies direct *Xist*-interacting proteins. **a**, A schematic overview of the RAP-MS method. **b**, The SILAC ratio (*Xist*/U1) for each *Xist*-enriched protein identified by RAP-MS for one representative sample of four biological replicates. For SHARP and RBM15, the enrichment values are indicated above their bars. **c**, Each *Xist*-interacting protein is shown (scaled to protein length). The locations of functional domains are shown.

These results demonstrate that the RAP-MS method identifies the majority of known RNA-interacting proteins, and that the proteins identified by RAP-MS are highly specific for the purified ncRNA complex.

To define the proteins that interact with *Xist* during the initiation of XCI, we UV-crosslinked SILAC-labelled mouse embryonic stem (ES) cells after *Xist* induction[13] and purified *Xist* from nuclear extracts (Methods). To control for background proteins or non-specific proteins that might interact with any nuclear RNA, we separately purified the abundant U1 small nuclear RNA, which is not expected to interact with the same proteins as *Xist*. We identified the proteins in each sample using liquid chromatography-mass spectrometry and calculated a SILAC ratio for each protein based on the intensity of all heavy or light peptides originating from the *Xist* or *U1* purification (Fig. 1a, Methods).

We identified 10 proteins that were enriched for *Xist* relative to *U1* (SILAC ratio >threefold, Fig. 1b). All 10 proteins were reproducibly enriched in multiple *Xist* purifications from independent biological samples (Methods). Consistent with the notion that these proteins are direct *Xist*-interacting proteins, 9 proteins contain well-characterized RNA binding domains (Fig. 1c).

The identified *Xist*-interacting proteins are SHARP, RBM15, MYEF2, CELF1, HNRNPC, LBR, SAF-A, RALY, HNRNPM, and PTBP1 (Fig. 1b). SAF-A (scaffold attachment factor-A, also known as HNRNPU) was previously shown to interact directly with *Xist* and is required for tethering *Xist* to the inactive X chromosome in differentiated cells[10]. In addition, 5 of these proteins have been previously implicated in transcriptional repression, chromatin regulation, and nuclear organization. These include SHARP (SMRT and HDAC associated repressor protein, also known as SPEN), a member of the SPEN family of transcriptional repressors, which directly interacts with the SMRT component (also known as NCOR2) of the nuclear co-repressor complex[16] that is known to interact with and activate HDAC3 deacetylation activity on chromatin[7] (Fig. 1c). Interestingly, we also identified RBM15, another member of the SPEN family of transcriptional repressors, which shares the same domain structure as SHARP, but seems to have a distinct functional role during development[17]. MYEF2 has been shown to function as a negative regulator of transcription in multiple cell types, although its mechanism of regulation is still unknown[18]. HNRNPM is a paralogue of MYEF2. Finally, we identified LBR (lamin B receptor), a protein that is anchored in the inner nuclear membrane and interacts with repressive chromatin regulatory proteins and lamin B[19] (Fig. 1c).

We confirmed the specificity of the identified *Xist*-interacting proteins (Supplementary Note 2, Methods). To ensure that they were not identified owing to non-specific RNA or protein capture, we performed RAP in uninduced cells (no *Xist*) and identified no enriched proteins. To ensure that these proteins are crosslinked with *Xist* in cells and not merely associating in solution, we performed RAP in cells that were not crosslinked (no UV) and identified no enriched proteins. To ensure that these proteins do not merely interact with any nuclear-enriched long ncRNA, we compared the *Xist*-purified proteins to those purified with 45S (pre-ribosomal RNA) and found that all 10 *Xist*-interacting proteins were still enriched. Finally to validate these interactions independently, we obtained high-quality affinity reagents for 8 of the 10 proteins (PTBP1, HNRNPC, CELF1, MYEF2, RBM15, LBR, RALY and SHARP), and immunoprecipitated the identified proteins in UV-crosslinked lysates. In all cases, we observed a strong enrichment for the *Xist* RNA (>fourfold), but not control mRNAs or lncRNAs (Extended Data Fig. 3, Supplementary Table 1).

Together, these results identify a set of highly specific and reproducible proteins that interact directly with *Xist* during the initiation of XCI. Given the generality of the RAP-MS approach, we expect that it will be broadly applicable for defining the proteins that interact directly with other lncRNAs.

To determine which proteins are required for *Xist*-mediated transcriptional silencing, we knocked down each of the proteins identified and assayed for the failure to silence gene expression on the X chromosome upon induction of *Xist* expression (Fig. 2a).

Specifically, we selected two X-linked genes, *Gpc4* and *Atrx*, that are well expressed in the absence of *Xist* expression, but are normally silenced by 16 h of *Xist* induction in our doxycycline-inducible system

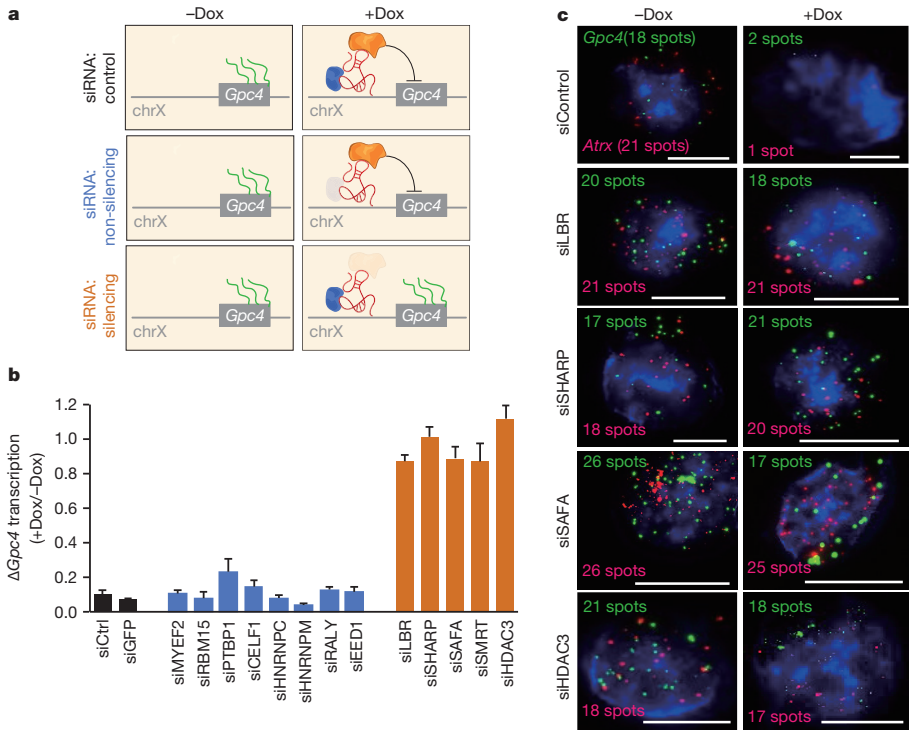


Figure 2 | SHARP, LBR and SAF-A are required for *Xist*-mediated gene silencing. **a**, Screen for *Xist*-mediated gene silencing for knockdown of control (top), non-silencing proteins (middle), or silencing proteins (bottom). **b**, *Gpc4* mRNA levels after induction of *Xist* (+Dox) normalized to *Gpc4* levels before *Xist* induction (−Dox). Error bars, standard error of the mean across 50 cells from one experiment. siCtrl, scrambled siRNA control. **c**, Images of individual cells for two X-linked mRNAs, *Gpc4* (green) and *Atrx* (red), and DAPI (blue) after treatment with different siRNAs (rows). The number of identified mRNAs is shown. Scale bars, 5 μm.

in male cells (Fig. 2b, Methods). We used short interfering RNAs (siRNAs) to knockdown the messenger RNA levels of each of the proteins identified by RAP-MS along with several negative controls (Methods, Supplementary Table 2). We ensured that each cell examined showed both successful depletion of the siRNA-targeted mRNA (>70% reduction) as well as induction of *Xist* expression using single-molecule RNA fluorescence *in situ* hybridization (FISH; Methods). Within each of these cells, we quantified the mRNA level of each of the two X-linked genes before *Xist* induction (−Dox) and after *Xist* induction (+Dox).

As a control, we transfected several non-targeting siRNAs (Methods). In these negative controls, we observed the expected silencing of the X-linked genes studied (*Gpc4* transcript levels decreased from an average of 20 copies (−Dox) to 2 copies (+Dox) per cell and *Atrx* transcript levels decreased from 22 to 3 copies per cell; Fig. 2b, c). Consistent with previous observations, we found no effect on X chromosome gene silencing upon knockdown of EED[20,21], a required component of PRC2[22] (Fig. 2b), or other proteins previously associated with *Xist* that do not seem to be required for transcriptional silencing (Extended Data Fig. 4, Methods). Similarly, knockdown of RBM15, MYEF2, PTBP1, CELF1, HNRNPC, RALY or HNRNPM did not alter gene silencing on the X chromosome (Fig. 2b, Extended Data Fig. 5).

In contrast, knockdown of SHARP, LBR or SAF-A largely abolished the silencing of X chromosome genes following *Xist* induction (Fig. 2b, c, Supplementary Note 3, Extended Data Figs 5, 6). Indeed, the expression levels of the X chromosome genes studied did not significantly change following *Xist* expression (Fig. 2c, Extended Data Fig. 5). These same silencing defects were observed with several independent siRNAs (Extended Data Fig. 7). Importantly, we observed the same X chromosome silencing defects upon knockdown of SHARP, LBR or SAF-A in differentiating female ES cells (Extended Data Fig. 8, Methods).

These results demonstrate that SHARP, LBR and SAF-A are required for *Xist*-mediated transcriptional silencing of the X chromosome. Although the remaining seven *Xist*-interacting proteins showed no effect on X-chromosome gene silencing, they may still be important for *Xist* function. Some may have redundant functions (for example, MYEF2 and HNRNPM, which are known paralogues), in some of these cases, the small amount of protein remaining after knockdown may still be sufficient for *Xist* function, or some of these proteins may be important for alternative *Xist*-mediated roles, such as the maintenance of XCI, which would not be captured by this silencing assay.

Xist initiates XCI by spreading across the future inactive X chromosome, excluding RNA polymerase II (Pol II), and repositioning active genes into a transcriptionally silenced nuclear compartment[3,10,13,23]. All of these roles—localization, RNA Pol II exclusion and repositioning—are required for proper silencing of transcription during the initiation of XCI[3].

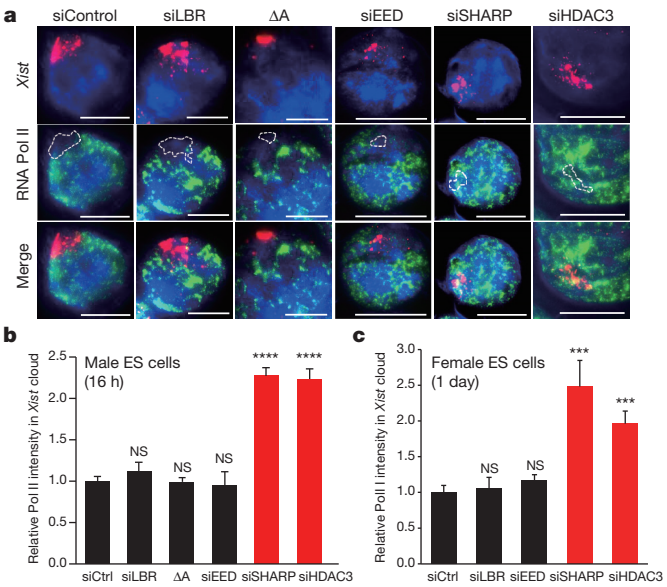


Figure 3 | SHARP is required for exclusion of Pol II from the *Xist*-coated territory. **a**, *Xist* (red), Pol II (green) and DAPI (blue) across different siRNA conditions (rows). **b**, **c**, Quantification of fluorescence intensity of Pol II within *Xist* territory normalized to control siRNA levels for male ES cells after 16 h of doxycycline treatment (**b**) and female ES cells after 1 day of retinoic-acid-induced differentiation (**c**). Error bars, standard error of the mean across 50 cells from one experiment. NS, not significant; ***P value < 0.005, ****P value < 0.001 relative to siControl by unpaired two-sample t-test. ΔA, genetic deletion of A-repeat of *Xist*. Scale bars, 5 μm.

Consistent with previous observations that SAF-A is required for *Xist* localization to chromatin in differentiated cells[10], we observed a diffuse *Xist* localization pattern in the nucleus upon knockdown of SAF-A (Extended Data Fig. 5). This suggests that SAF-A is required for transcriptional silencing by localizing *Xist*, and its silencing proteins, to the X chromosome during the initiation of XCI.

To determine the proteins responsible for establishing the initial silenced compartment on the X chromosome, we explored whether SHARP or LBR are required for the exclusion of Pol II from the *Xist*-coated region. Specifically, we measured the co-localization of *Xist* and Pol II in single cells (Methods). In wild-type cells after 16 h of *Xist* induction, we observed a depletion of Pol II over the *Xist*-coated territory (Fig. 3a, Methods). We observed a similar exclusion of Pol II from the *Xist*-coated region in the negative controls and upon knockdown of EED or LBR (Fig. 3a, b). In contrast, upon knockdown of SHARP, we observed higher levels of Pol II over the *Xist*-coated territory relative to the control samples (Fig. 3b). We confirmed that SHARP, but not LBR or EED, is similarly required for Pol II exclusion in differentiating female ES cells (Fig. 3c, Extended Data Fig. 9, Supplementary Note 4).

These results demonstrate that SHARP is required to exclude Pol II on the inactive X chromosome and may be required for creating the initial silenced compartment upon *Xist* localization[3]. Although LBR is not required for Pol II exclusion, it is likely to have an alternative role during the initiation of *Xist*-mediated transcriptional silencing, such as repositioning genes into this Pol II-excluded compartment[13,23].

Having identified SHARP as the direct *Xist*-interacting protein that is required for excluding Pol II on the X chromosome, we sought to determine how it might carry out this role. SHARP is a direct RNA binding protein[6,24] that was first identified in mammals on the basis of its interaction with the SMRT co-repressor complex[6], which is known to interact with HDAC3 and is required for activating its deacetylation and transcriptional silencing activity *in vivo*[7]. Based on these previous observations, we hypothesized that *Xist*-mediated transcriptional silencing through SHARP would occur through SMRT and the silencing function of HDAC3. (We would not expect to identify these proteins by RAP-MS, which was designed to identify only direct RNA–protein interactions.)

To test this hypothesis, we knocked down either SMRT or HDAC3 and measured the expression of X chromosome genes upon *Xist* induction. Knockdown of SMRT or HDAC3 in both male and female ES cells abrogated silencing of X chromosome genes upon induction of *Xist* expression (Fig. 2b, Extended Data Figs 5, 7 and 8). To ensure that the observed silencing defect is specific for HDAC3 and not for other class I HDAC proteins, we knocked down HDAC1 or HDAC2 and observed no effect on gene silencing (Extended Data Fig. 5). To further confirm the specificity of our results, we used independent siRNAs to knockdown SMRT or HDAC3 and in all cases identified a similar silencing defect (Extended Data Fig. 7).

To determine whether this effect is similar to the effect produced by knockdown of SHARP or a distinct defect in transcriptional silencing, we tested whether HDAC3, the silencing protein in this complex[7,25], is required for the exclusion of RNA Pol II from the *Xist*-coated territory. We found that knockdown of HDAC3 in both male and female ES cells eliminated the exclusion of RNA Pol II from the *Xist*-coated compartment to a similar degree to that seen for knockdown of SHARP (Fig. 3, Extended Data Fig. 9).

These results suggest that SHARP silences transcription through SMRT and the HDAC3 silencing protein. This role for HDAC3 in *Xist*-mediated silencing would explain the long-standing observation of global hypoacetylation on the entire X chromosome as one of the very first events that occur upon initiation of XCI[3,26].

One of the features of XCI is the recruitment of PRC2 and its associated H3K27me3 repressive chromatin modifications across the X chromosome in an *Xist*-dependent manner[3,4,9]. Although PRC2 is not required for the initiation of XCI[11,12] (Fig. 3b), it or its associated

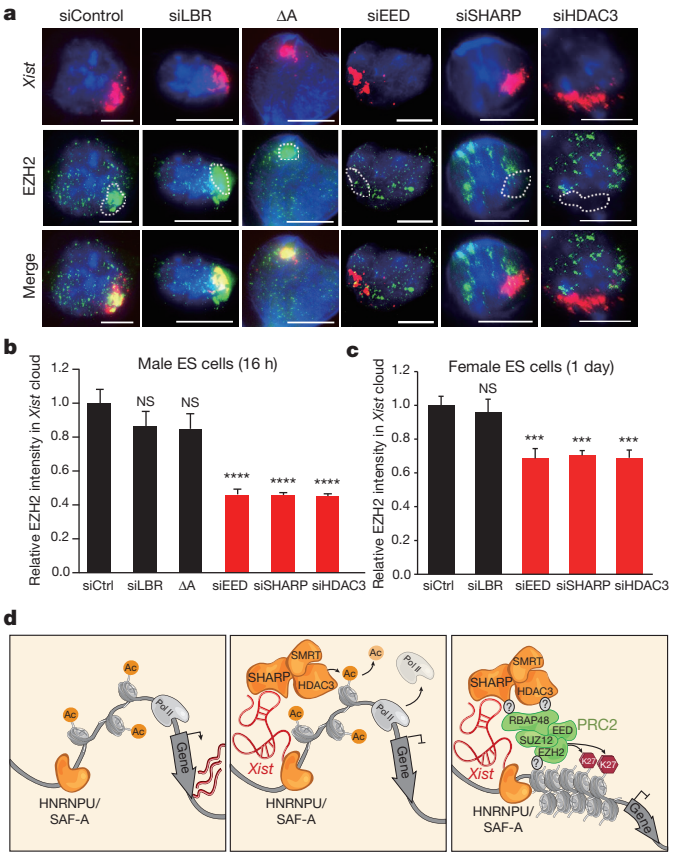


Figure 4 | SHARP is required for PRC2 recruitment across the *Xist*-coated territory. **a**, *Xist* (red), EZH2 (green) and DAPI (blue) across siRNA conditions (rows). **b**, **c**, Quantification of EZH2 levels within the defined *Xist* territory normalized to the levels in the control siRNA sample for male ES cells (**b**) and differentiating female ES cells (**c**). Error bars, standard error of the mean across 50 cells from one experiment. NS, not significant; ***P value < 0.005, ****P value < 0.001 relative to siControl by an unpaired two-sample t-test. Scale bars, 5 μm. **d**, A model for *Xist*-mediated transcriptional silencing and recruitment of PRC2 across the X chromosome.

H3K27me3 repressive chromatin modifications may be involved in establishing an epigenetically silenced state[27]. Yet, how *Xist* recruits the PRC2 complex across the X chromosome is unknown. Since we failed to identify any PRC2 components by RAP-MS, and various HDAC complexes are known to recruit PRC2[28], we hypothesized that PRC2-recruitment is mediated by SHARP and HDAC3.

To test this hypothesis, we looked at PRC2 recruitment to the *Xist*-coated territory. In wild-type cells, we observe a strong enrichment of EZH2, a component of PRC2, over the *Xist*-coated territory after 16 h of induction (Fig. 4a). Upon knockdown of EED, a distinct component of the PRC2 complex that is required for its proper localization to chromatin[22], we observe no enrichment of EZH2 over the *Xist*-coated territory at this same time point (Fig. 4a). Similarly, upon knockdown of SHARP, we identified a loss of EZH2 over the *Xist*-coated territory, of comparable magnitude to that observed in the absence of EED (Fig. 4a). Conversely, upon knockdown of LBR, we observed a strong enrichment of EZH2 over the *Xist*-coated territory, of comparable magnitude to the levels of recruitment in wild-type conditions (Fig. 4b). To determine whether HDAC3 is required for PRC2 recruitment, we knocked down HDAC3 and observed a loss of PRC2 recruitment (Fig. 4a), of comparable magnitude to that observed upon loss of SHARP (Fig. 4b). Knockdown of SHARP or HDAC3 led to the same PRC2-recruitment defect in female ES cells (Fig. 4c, Extended Data Fig. 10).

These results indicate that *Xist*-mediated recruitment of PRC2 across the X chromosome is dependent on SHARP and HDAC3. Whether this occurs through an interaction with SHARP or HDAC3

(direct recruitment) or due to the HDAC3-induced silenced transcription state, chromatin modifications, or compact chromatin structure (indirect recruitment) remains unclear (Supplementary Note 5). Yet, our results are in contrast to a previous model that PRC2 is recruited through a direct interaction between EZH2 and the A-repeat of *Xist*[8]. The evidence for this PRC2–*Xist* interaction is based on *in vitro* binding and purifications in non-denaturing conditions[8]. Recently, the specificity of this interaction has been questioned because PRC2 appears to bind promiscuously to many RNAs, including bacterial RNAs, in these conditions[29]. Instead, our results are consistent with reports that deletion of the A-repeat, unlike knockdown of SHARP or HDAC3, has no significant effect on PRC2 recruitment to the *Xist*-coated territory[9] (Fig. 4b).

Taken together, our data suggest a model for how *Xist* can orchestrate transcriptional silencing on the X chromosome (Fig. 4d). Upon initiation of *Xist* expression, *Xist* can localize to sites on the X chromosome by binding to the SAF-A protein[10], which is known to interact directly with chromatin[30]. *Xist* interacts directly with SHARP to recruit SMRT[6] to these DNA sites across the inactive X chromosome. This *Xist*–SHARP–SMRT complex either recruits HDAC3 directly to the X chromosome or may act to induce the enzymatic activity of HDAC3[7] that may already be present at active genes across the X chromosome[31]. Through HDAC3, *Xist* can direct the removal of activating histone acetylation marks on chromatin, thereby compacting chromatin and silencing transcription[32]. Upon initiating the silenced state, *Xist* recruits PRC2 across the X chromosome in an HDAC3-dependent manner, either through a direct interaction between PRC2 and HDAC3 or indirectly through HDAC3-induced transcriptional silencing or chromatin compaction (Supplementary Note 5). In this way, the same *Xist*-interacting protein might achieve two essential roles in XCI: initiating the inactive state by recruiting transcriptional silencers (HDAC3) and maintaining the inactive state by recruiting stable epigenetic silencers (PRC2)[27]. Beyond *Xist*, RAP-MS provides a critical tool that will accelerate the discovery of novel lncRNA mechanisms that have thus far proved elusive.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 November 2014; accepted 2 April 2015.
Published online 27 April 2015.

1. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477,** 295–300 (2011).
2. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
3. Wutz, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Rev. Genet.* **12,** 542–553 (2011).
4. Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.* **23,** 1831–1842 (2009).
5. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol.* **15,** 203 (2014).
6. Shi, Y. *et al.* Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. *Genes Dev.* **15,** 1140–1151 (2001).
7. You, S. H. *et al.* Nuclear receptor co-repressors are required for the histone-deacetylase activity of HDAC3 *in vivo*. *Nature Struct. Mol. Biol.* **20,** 182–187 (2013).
8. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
9. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300,** 131–135 (2003).
10. Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K. & Nakagawa, S. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* **19,** 469–476 (2010).
11. Schoeftner, S. *et al.* Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J.* **25,** 3110–3122 (2006).
12. Kalantry, S. & Magnuson, T. The Polycomb group protein EED is dispensable for the initiation of random X-chromosome inactivation. *PLoS Genet.* **2,** e66 (2006).
13. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341,** 1237973 (2013).
14. Darnell, R. B. HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdiscipl. Rev. RNA* **1,** 266–286 (2010).
15. Ong, S. E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protoc.* **1,** 2650–2660 (2007).
16. Ariyoshi, M. & Schwabe, J. W. A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling. *Genes Dev.* **17,** 1909–1920 (2003).
17. Raffel, G. D. *et al.* Ott1(Rbm15) has pleiotropic roles in hematopoietic development. *Proc. Natl Acad. Sci. USA* **104,** 6001–6006 (2007).
18. Haas, S., Steplewski, A., Siracusa, L. D., Amini, S. & Khalili, K. Identification of a sequence-specific single-stranded DNA binding protein that suppresses transcription of the mouse myelin basic protein gene. *J. Biol. Chem.* **270,** 12503–12510 (1995).
19. Olins, A. L., Rhodes, G., Welch, D. B., Zwerger, M. & Olins, D. E. Lamin B receptor: multi-tasking at the nuclear envelope. *Nucleus* **1,** 53–70 (2010).
20. Brown, C. J. & Baldry, S. E. Evidence that heteronuclear proteins interact with XIST RNA *in vitro*. *Somat. Cell Mol. Genet.* **22,** 403–417 (1996).
21. Sarma, K. *et al.* ATRX directs binding of PRC2 to Xist RNA and Polycomb targets. *Cell* **159,** 869–883 (2014).
22. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–349 (2011).
23. Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20,** 2223–2237 (2006).
24. Arieti, F. *et al.* The crystal structure of the Split End protein SHARP adds a new layer of complexity to proteins containing RNA recognition motifs. *Nucleic Acids Res.* **42,** 6742–6752 (2014).
25. Li, J., Lin, Q., Wang, W., Wade, P. & Wong, J. Specific targeting and constitutive association of histone deacetylase complexes during transcriptional repression. *Genes Dev.* **16,** 687–692 (2002).
26. Keohane, A. M., O'Neill, L. P., Belyaev, N. D., Lavender, J. S. & Turner, B. M. X-inactivation and histone H4 acetylation in embryonic stem cells. *Dev. Biol.* **180,** 618–630 (1996).
27. Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol. Cell* **55,** 347–360 (2014).
28. van der Vlag, J. & Otte, A. P. Transcriptional repression mediated by the human polycomb-group protein EED involves histone deacetylation. *Nature Genet.* **23,** 474–478 (1999).
29. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature Struct. Mol. Biol.* **20,** 1250–1257 (2013).
30. Fackelmayer, F. O., Dahm, K., Renz, A., Ramsperger, U. & Richter, A. Nucleic-acid-binding properties of hnRNP-U/SAF-A, a nuclear-matrix protein which binds DNA and RNA *in vivo* and *in vitro*. *Eur. J. Biochem.* **221,** 749–757 (1994).
31. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138,** 1019–1031 (2009).
32. Kuo, M. H. & Allis, C. D. Roles of histone acetyltransferases and deacetylases in gene regulation. *Bioessays* **20,** 615–626 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Engreitz for extensive discussions, help in adapting the RAP method, and critical comments on the manuscript; A. Gnirke, S. Carr, J. Jaffe and M. Schenone for initial discussions about the RAP-MS method; A. Collazo, E. Lubek, and L. Cai for microscopy help; A. Wutz for providing transgenic cell lines; R. Eggleston-Rangel for assistance with mass spectrometry; S. Grossman, I. Amit, M. Garber and J. Rinn for comments on the manuscript and helpful suggestions; and S. Knemeyer for illustrations. C.A.M. is supported by a post-doctoral fellowship from Caltech. C.-K.C. is supported by an NIH NRSA training grant (T32GM07616). Imaging was performed in the Biological Imaging Facility, with the support of the Caltech Beckman Institute and the Arnold and Mabel Beckman Foundation. This work was funded by the Gordon and Betty Moore Foundation (GBMF775), the Beckman Institute, and NIH (1S10RR029591-01A1 to S.H.), an NIH Director's Early Independence Award (DP5OD012190), the Rose Hills Foundation, Edward Mallinckrodt Foundation, Sontag Foundation, Searle Scholars Program, and funds from the California Institute of Technology.

Author Contributions C.A.M. developed the RAP-MS method, designed, performed, and analysed RAP-MS experiments and data, C.-K.C. designed, performed, and analysed *Xist* functional experiments, A.C. designed, performed, and oversaw experiments, C.F.S. helped develop RAP-MS and performed experiments, C.T., P.M., A.P.-J., A.M., A.A.S., J.S. performed experiments, M.J.S., M.B., C.B. analysed data, E.S.L. helped develop initial ideas for adapting RAP for protein detection, S.H. oversaw mass spectrometry development and data analysis, K.P. helped design *Xist* RAP-MS and functional experiments and analysed data, M.G. conceived, designed and oversaw the entire project and integrated the data, C.A.M., C.-K.C. and M.G. wrote the manuscript with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G. (mguttman@caltech.edu).

METHODS

No statistical methods were used to predetermine sample size. Detailed RAP-MS protocols are available at the authors' web site: (http://www.lncrna.caltech.edu/RAP).

Mouse ES cell culture. All mouse ES cell lines were cultured in serum-free 2i/LIF medium as previously described[13]. We used the following cell lines: wild-type male ES cells (V6.5 line); male ES cells expressing *Xist* from the endogenous locus under control of a tet-inducible promoter (pSM33 ES cell line) as previously described[13]; male ES cells carrying a cDNA *Xist* transgene without the A-repeat integrated into the *Hprt* locus under control of the tet-inducible promoter (A-repeat deletion: provided by A. Wutz)[33]; female ES cells (F1 2-1 line). This wild-type female mouse ES cell line is derived from a 129 × castaneous F1 mouse cross as previously described[13].

***Xist* induction.** For Dox inducible cells (pSM33 and A-repeat deletion), we induced *Xist* expression by treating cells with 2 μg ml^{-1} doxycycline (Sigma) for 6 h, 16 h, or 24 h based on the application. For female ES cells (F1 2-1 line), we induced *Xist* expression by inducing differentiation; 2i was replaced with MEF media (DMEM, 10% Gemini Benchmark FBS, 1 × L-glutamine, 1 × NEAA, 1 × penicillin/streptomycin; Life Technologies unless otherwise indicated) for 24 h followed by treatment with 1 μM retinoic acid (RA) (Sigma) for an additional 24 h.

We measured the amount of *Xist* RNA in both the doxycycline-inducible cells (6 h induction) and differentiating female ES cells (24 h induction) by qRT–PCR. We normalized this level to various RNA housekeeping controls, 18S, 28S, and U6, in both cell populations and calculated the fold expression difference between male and female cells using the comparative Ct method. We observed a range of expression, with the male inducible system expressing from 5 to 20-fold (12-fold average) more *Xist* than the female cells. We note that this estimate likely represents an upper limit of the actual differences because the female ES cell system is known to be heterogeneous in *Xist*-induction, such that not every cell will induce *Xist* to the same level after 24 h of retinoic acid treatment. Accordingly, we expect that the actual differences between the male inducible system and differentiating female ES cells are actually significantly lower. While the precise levels are hard to compare by single molecule FISH, the size and intensity of each *Xist* RNA-coated territory is similar in both systems at the time points used.

The male-inducible system is more sensitive for identifying proteins that affect silencing compared to a female system because *Xist*-mediated silencing in males will lead to loss of 100% of X-chromosome transcripts rather than only 50% in a female system, which still retains one active X.

UV crosslinking. Cells were washed once with PBS and then crosslinked on ice using 0.8 J cm^{-2} (UV8k) of UV at 254 nm in a Spectrolinker UV Crosslinker. Cells were then scraped from culture dishes, washed once with PBS, pelleted by centrifugation at 1,500*g* for 4 min, and flash-frozen in liquid nitrogen for storage at –80 °C.

SILAC ES cell culture. For SILAC experiments, we adapted our ES cell culture procedures to incorporate either light or heavy lysine and arginine amino acids. The 2i/LIF SILAC medium was composed as follows: custom DMEM/F-12 without lysine or arginine (Dundee Cell Products) was supplemented with either 0.398 mM heavy Arg10 (Sigma) or unlabelled arginine (Sigma) and either 0.798 mM heavy Lys8 (Cambridge Isotope Labs) or unlabelled lysine (Sigma), 0.5 × B-27 (Gibco), 2 mg ml^{-1} bovine insulin (Sigma), 1.37 μg ml^{-1} progesterone (Sigma), 5 mg ml^{-1} BSA Fraction V (Gibco), 0.1 mM 2-mercaptoethanol (Sigma), 5 ng ml^{-1} murine LIF (GlobalStem), 0.1 μM PD0325901 (SelleckChem) and 0.3 μM CHIR99021 (SelleckChem). Cells in both heavy and light 2i/LIF SILAC medium were also supplemented with 0.2 mg ml^{-1} of unlabelled proline (Sigma) to prevent conversion of labelled arginine to proline. 2i inhibitors were added fresh with each medium change.

Adapting cells to SILAC conditions. Prior to mass spectrometry, ES cells were adapted to SILAC conditions over three passages. The heavy or light culture medium was replaced every 24–48 h depending on cell density, and cells were passaged every 72 h using 0.025% trypsin (Gibco), rinsing dissociated cells from the plates with DMEM/F12 containing 0.038% BSA Fraction V (Gibco). Cells were grown in two different types of medium: 2i/LIF SILAC medium with light (unlabelled) lysine and arginine, or 2i/LIF SILAC medium with heavy isotope-labelled lysine and arginine.

Measuring SILAC incorporation. To examine the efficiency of SILAC labelling in pSM33 cells, we tested for the incorporation of labelled amino acids after 10 days of growth (3 cell passages) in heavy 2i/LIF SILAC medium. Pellets of 2 million cells were boiled for 10 min in LDS Sample Loading Buffer (Invitrogen) and then proteins were separated by SDS–PAGE on a 4–12% Tris-Glycine polyacrylamide gel (Invitrogen). Total protein was stained with Colloidal Coomassie (Invitrogen) and gel slices were excised with a clean scalpel and transferred to microcentrifuge tubes for in-gel tryptic digest. Protein disulphide bonds were reduced with DTT then alkylated with iodoacetamide. Proteins were digested with trypsin overnight and then extracted using successive washes with 1% formic acid/2% acetonitrile, 1:1 acetonitrile/water, and 1% formic acid in acetonitrile. Peptides were collected, lyophilized, then resuspended in 1% formic acid for mass spectrometry analysis (described later in Mass spectrum measurements). Peptides were identified from mass spectra using MaxQuant (described later in Mass spectrometry data analysis). The incorporation rate of labelled amino acids was calculated based on the ratio of the intensity of heavy and light versions of each peptide identified. In cells used for subsequent assays, we confirmed that over 95% of peptides from cellular proteins showed >95% incorporation of labelled amino acids (Extended Data Fig. 1b).

RNA antisense purification-mass spectrometry (RAP-MS). *Probe design and generation.* To create the probes used to capture target RNAs, we designed and synthesized 90-mer DNA oligonucleotides (Eurofins Operon) that spanned the entire length of the target RNA. The sequence of each DNA oligonucleotide probe was antisense to the complementary target RNA sequence. Each DNA oligonucleotide probe was also modified with a 5′ biotin in order to enable capture of DNA-RNA hybrids on streptavidin coated magnetic beads (described below). While we had previously used 120-mer probes, we found that 90-mer probes provided comparable stringency and yield in the conditions used. For *Xist*, we used 142 probes that covered the entire mature RNA sequence, with the exception of regions that match to other transcripts or genomic regions as previously described[13,34].

Total cell lysate preparation. For the 18S and U1 experiments we used total cellular lysates prepared in the following manner. We lysed batches of 20 million cells by completely resuspending frozen cell pellets in ice cold detergent-based Cell Lysis Buffer (10 mM Tris pH 7.5, 500 mM LiCl, 0.5% dodecyl maltoside (DDM, Sigma), 0.2% sodium dodecyl sulphate (SDS, Ambion), 0.1% sodium deoxycholate (Sigma)). Next, 1 × Protease Inhibitor Cocktail (Set III, EDTA-free, Calbiochem) and 920 U of Murine RNase Inhibitor (New England Biolabs) were added and the sample was incubated for 10 min on ice to allow lysis to proceed. During this incubation period, the cell sample was passed 3–5 times through a 26-gauge needle attached to a 1 ml syringe in order to disrupt the pellet and shear genomic DNA. Each sample was then sonicated using a Branson Digital Sonifier with a microtip set at 5 W power for a total of 30 s in intermittent pulses (0.7 s on, 1.3 s off). During sonication the samples were chilled to prevent overheating of the lysate. The samples were then treated for 10 min at 37 °C with 2.5 mM $MgCl_2$, 0.5 mM $CaCl_2$, and 20 U of TURBO DNase (Ambion) to digest DNA. Samples were returned to ice and the reaction was immediately terminated by the addition of 10 mM EDTA and 5 mM EGTA. Disulphide bonds were reduced by addition of 2.5 mM Tris-(2-carboxyethyl) phosphine (TCEP) and samples were then mixed with twice the lysate volume of 1.5 × LiCl/Urea Buffer (the final1 × buffer contains 10 mM Tris pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% deoxycholate, 4 M urea, 2.5 mM TCEP). Lysates were incubated on ice for 10 min then cleared by centrifugation in an Eppendorf 5424R centrifuge for 10 min at 16,000*g*. Supernatants were pooled and flash frozen in liquid nitrogen for storage at −80 °C.

Nuclear lysate preparation. For the *Xist* versus U1 and 45S versus U1 comparisons, we used nuclear lysates prepared in the following manner. We lysed batches of 50 million cells by resuspending frozen pellets in 1 ml Lysis Buffer 1 (10 mM HEPES pH 7.2, 20 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM EDTA, 1 mM Tris(2-carboxyethyl)-phosphine (TCEP), 0.5 mM PMSF). Then the samples were centrifuged at 3,300*g* for 10 min to pellet cells. The cell pellets were resuspended in 1 ml Lysis Buffer 1 with 0.1% dodecyl maltoside (DDM) and Dounced 20 times using a glass Dounce homogenizer with the small clearance pestle (Kontes). Nuclei released from the cells after Douncing were pelleted by centrifugation at 3,300*g* then resuspended in 550 μl Lysis Buffer 2 (20 mM Tris pH 7.5, 50 mM KCl, 1.5 mM $MgCl_2$, 2 mM TCEP, 0.5 mM PMSF, 0.4% sodium deoxycholate, 1% DDM, and 0.1% N-lauroylsarcosine (NLS)). Samples were incubated on ice for 10 min, then each sample was sonicated using a Branson Sonifier at 5 W power for a total of 1 min in intermittent pulses (0.7 s on, 3.3 s off) to lyse nuclei and solubilize chromatin. During sonication the samples were chilled to prevent overheating of the nuclear lysate. Samples were then treated with 2.5 mM $MgCl_2$, 0.5 mM $CaCl_2$, and 330 U TURBO DNase (Ambion) for 12 min at 37 °C to further solubilize chromatin. After DNase treatment, lysates were mixed with equal volume of 2 × Hybridization Buffer (the final 1 × Buffer contains 10 mM Tris pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% deoxycholate, 4 M urea, 2.5 mM TCEP). Finally, lysates were cleared by centrifugation for 10 min at 16,000g in an Eppendorf 5424R centrifuge and the resulting supernatants were pooled and flash frozen in liquid nitrogen for storage at −80 °C.

RNA antisense purification of crosslinked complexes. Lysates from 200 million or 800 million cells were used for each capture. For 200 million cells the following protocol was used, and scaled appropriately for larger cell numbers. For each capture, a sample of heavy or light SILAC labelled frozen lysate was warmed to 37 °C. For each sample, 1.2 ml of Streptavidin Dynabeads MyOne C1 magnetic

beads (Invitrogen) were washed 6 times with equal volume of hybridization buffer (10 mM Tris pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% deoxycholate, 4 M urea, 2.5 mM TCEP). Lysate samples were pre-cleared by incubation with the washed Streptavidin C1 magnetic beads at 37 °C for 30 min with intermittent shaking at 1,100 r.p.m. on a Eppendorf Thermomixer C (30 s mixing, 30 s off). Streptavidin beads were then magnetically separated from lysate samples using a Dynamag magnet (Life Technologies). The beads used for pre-clearing lysate were discarded and the lysate sample was transferred to fresh tubes twice to remove all traces of magnetic beads. Biotinylated 90-mer DNA oligonucleotide probes specific for the RNA target of interest (20 µg per sample, in water) were heat-denatured at 85 °C for 3 min and then snap-cooled on ice. Probes and pre-cleared lysate were mixed and incubated at 67 °C using an Eppendorf thermomixer with intermittent shaking (30 s shaking, 30 s off) for 2 h to hybridize probes to the capture target RNA. Hybrids of biotinylated DNA probes and target RNA were then bound to streptavidin beads by incubating each sample with 1.2 ml of washed Streptavidin coated magnetic beads at 67 °C for 30 min on an Eppendorf Thermomixer C with intermittent shaking as above. Beads with captured hybrids were washed 6 times with LiCl/Urea Hybridization Buffer at 67 °C for 5 min to remove non-specifically associated proteins. Between 0.5 and 1% of the total beads were removed and transferred to a fresh tube after the final wash to examine RNA captures by qPCR (see Elution and analysis of RNA samples). The remaining beads were resuspended in Benzonase Elution Buffer (20 mM Tris pH 8.0, 2 mM $MgCl_2$, 0.05% NLS, 0.5 mM TCEP) for subsequent processing of the protein samples.

Elution of protein samples. Elution of captured proteins from streptavidin beads was achieved by digesting all nucleic acids (both RNA and DNA, double-stranded and single-stranded) using 125 U of Benzonase nonspecific RNA/DNA nuclease for 2 h at 37 °C (Millipore, #71206-3). Beads were then magnetically separated from the sample using a DynaMag magnet (Life Technologies) and the supernatant containing eluted *Xist*-specific proteins were precipitated overnight at 4 °C with 10% trichloroacetic acid (TCA). TCA treated protein elution samples were pelleted by centrifugation for 30 min at >20,000*g*, then washed with 1 ml cold acetone and recentrifuged. Final protein elution pellets were air dried to remove acetone and stored at −20 °C until processing for mass spectrometry.

Elution and analysis of RNA samples. Beads with hybrids were magnetically separated using a 96-well DynaMag (Life Technologies) and the supernatant was discarded. Beads were then resuspended by pipetting in 20 µl NLS RNA Elution Buffer (20 mM Tris pH 8.0, 10 mM EDTA, 2% NLS, 2.5 mM TCEP). To release the target RNA, beads were heated for 2 min at 95 °C in an Eppendorf Thermomixer C. Beads were then magnetically separated using a 96-well DynaMag (Life Technologies) and the supernatants containing eluted target RNA were digested by the addition of 1 mg ml^{-1} Proteinase K for 1 h at 55 °C to remove all proteins. The remaining nucleic acids were then purified by ethanol precipitation onto SILANE beads (Invitrogen) as previously described[13,34]. DNA probes were removed by digestion with TURBO DNase (Ambion). To quantify RNA yield and enrichment, qPCR was performed as previously described[13].

Mass spectrometry analysis. *Preparation of proteins for mass spectrometry.* Proteins from RAP-MS captures were resuspended in fresh 8 M urea dissolved in 40 µl of 100 mM Tris-HCl pH 8.5. Disulphide bonds were reduced by incubation with 3 mM TCEP for 20 min at room temperature, followed by alkylation with 11 mM iodoacetamide for 15 min at room temperature in the dark. Samples were then digested with 0.1 µg endoproteinase Lys-C for 4 h at room temperature. After Lys-C digestion the samples were diluted to a final concentration of 2 M urea by the addition of 100 mM Tris-HCl pH 8.5, and $CaCl_2$ was added to a final concentration of 1 mM. Tryptic peptides were generated by treatment with 0.1 to 0.5 µg of trypsin overnight at room temperature. Contaminating detergents were removed from peptides using HiPPR detergent removal columns (Thermo), and peptides were protonated by the addition of 5% formic acid before desalting on a Microm Bioresources C8 peptide MicroTrap column. Peptide fractions were collected and lyophilized, and dried peptides were resuspended in 0.2% formic acid with 5% acetonitrile.

Mass spectrum measurements. Liquid chromatography-mass spectrometry and data analyses of the digested samples were carried out as previously described[35] with the following modifications. All experiments were performed on a nanoflow LC system, EASY-nLC 1000 coupled to a hybrid linear ion trap Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoelectrospray ion source (Thermo Fisher Scientific). For the EASY-nLC II system, solvent A consisted of 97.8% H_2O, 2% acetonitrile, and 0.2% formic acid and solvent B consisted of 19.8% H_2O, 80% acetonitrile, and 0.2% formic acid. For the LC-MS/MS experiments, 200 ng of digested peptides were directly loaded at a flow rate of 500 nl min^{-1} onto a 16-cm analytical HPLC column (75 µm ID) packed in-house with ReproSil-Pur C_{18}AQ 3 µm resin (120 Å pore size, Dr. Maisch, Ammerbuch, Germany). The column was enclosed in a column heater operating at 30 °C. After 30 min of loading time, the peptides were separated with a 75 min gradient at a flow rate of 350 nl min^{-1}. The gradient was as follows: 0–2% Solvent B (5 min), 2–30% B (60 min), and 100% B (10 min). The Elite was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z = 400–1,600) in the Orbitrap and subsequent rapid 20 collision-induced dissociation (CID) MS/MS scans in the linear ion trap. CID was performed with helium as collision gas at a normalized collision energy of 35% and 10 ms of activation time.

MS data analysis. Thermo RAW files were searched with MaxQuant (v 1.5.0.30)[36,37]. Spectra were searched against all UniProt mouse entries (43,565 entries, downloaded 02 Oct 14) and MaxQuant contaminant database (245 entries). Decoy sequences (reversed peptide sequences) were generated in MaxQuant to estimate the false discovery rate[38]. Search parameters included multiplicity of 2 with heavy Arg (+10.0083) and heavy Lys (+8.0142) as heavy peptide modifications. Variable modifications included oxidation of Met (+15.9949) and protein N-terminal acetylation (+42.0106). Carboxyamidomethylation of Cys (+57.0215) was specified as a fixed modification. Protein and peptide false discovery rates were thresholded at 1%. Precursor mass tolerance was 7 p.p.m. (or less for individual peptides). Fragment mass tolerance was 0.5 Da. Requantify and match between runs were both enabled. Trypsin was specified as the digestion enzyme with up to 2 missed cleavages.

Identification of RNA-interacting proteins. Proteins of interest from RAP-MS captures were identified based on several criteria. First, proteins were considered identified only if 2 or more unique peptides were found in the mass spectrum. Then proteins of interest were selected based on the SILAC ratio of capture versus control samples. SILAC ratios for each peptide were calculated based on the intensity ratios of heavy and light SILAC pairs. The protein ratio is the median of all calculated peptide ratios, with a minimum of two SILAC pairs required for a SILAC ratio to be assigned to a given protein. A SILAC ratio cutoff of ≥ 3.0 (fold enrichment over control sample) was used as a cutoff for further analysis. We excluded known contaminants, including human keratin and proteins introduced during the sample purification and preparation process (such as streptavidin, benzonase, and trypsin), as well as naturally biotinylated proteins that contaminate the preparation by binding to streptavidin beads.

RAP-MS experiments and controls. *18S rRNA versus U1 snRNA.* To validate the RAP-MS method and identify proteins specifically interacting with 18S ribosomal RNA or U1 snRNA, we performed captures of each target RNA in parallel samples from heavy and light labelled lysates from wild-type V6.5 ES cells. The total protein quantity in elution samples from each RAP-MS capture was measured by comparing the median intensity of peptides identified in a single quantitation MS run for each sample. The heavy and light label swapped samples were then mixed equally based on total protein quantity and analysed by mass spectrometry to identify the SILAC enrichment ratio of proteins originating from 18S ribosomal RNA or U1 snRNA captures. The experiment was performed twice and each experimental set contained two biological replicates of 18S and U1 captures (heavy and light labelling states).

Xist lncRNA versus U1 snRNA captures. To identify proteins specifically interacting with *Xist* lncRNA, we performed captures as described above with either 200 million or 800 million pSM33 cells treated with doxycycline for 6 h. The total protein quantity in elution samples from each RAP-MS capture was measured by a single quantitation MS run for each sample. Heavy and light label swapped samples were mixed equally based on total protein quantity, and analysed by mass spectrometry. SILAC ratios of *Xist*-enriched proteins versus U1-enriched proteins were calculated and used to identify *Xist*-specific interacting proteins for further analysis. The experiment was performed twice and each experimental set contained two biological replicates of *Xist* and U1 captures, from heavy and light labelled samples. Proteins replicated well between samples, with a sole exception (LBR) that was missed only because its enrichment level (twofold) fell below our enrichment cutoff (threefold) in some replicate samples.

Xist lncRNA capture from non-crosslinked cells. As a control to ensure that purified proteins are not non-specifically associated or binding *in vitro* with target RNAs during capture, we performed RAP-MS captures of *Xist* from non-crosslinked cells otherwise treated in the same manner (that is, doxycycline treated for 6 h).

Xist lncRNA capture from cells where Xist is not expressed. To confirm that the identified proteins are not resulting from background proteins or probe association with other RNAs or proteins in the pSM33 cells, we performed RAP captures of *Xist* from pSM33 cells that were not treated with doxycycline, but which were otherwise treated identically.

45S pre-rRNA capture versus U1 capture. To ensure that the proteins enriched in *Xist* captures using RAP-MS are not simply due to increased protein capture as a consequence of long target RNA transcripts, we additionally performed captures of the 13,000 nucleotide long 45S pre-ribosomal RNA as a control. To ensure specific capture only of the 45S, and not the mature 18S and 28S, we designed

probes that specifically targeted the internal transcribed spacer regions (ITS1 and ITS2) that are only present in the 45S pre-ribosomal RNA. The experiment was performed in the same manner and with the same conditions as the *Xist* lncRNA captures described above. To compare *Xist* protein enrichment to 45S protein enrichment, we used a SILAC approach based on direct comparison of two samples that share a common denominator (called spike-in SILAC[39]). Specifically, we calculated an overall *Xist*/45S SILAC ratio by multiplying the *Xist*/U1 ratio by the U1/45S ratio for each identified protein.

Protein domain classification. We defined the conserved domain structures of proteins using the Protein Families database (Pfam[40]).

RNA immunoprecipitation in UV-crosslinked cells. We crosslinked pSM33 cells after 6 h of doxycycline-treatment with 0.4 J cm^{-2} of $UV_{254\,nm}$. Cells were lysed and RNA was digested with RNase I to achieve a size range of 100–500 nucleotides in length. Lysate preparations were precleared by mixing with Protein G beads for 1 h at 4 °C. For each sample, target proteins were immunoprecipitated from 20 million cells with 10 μg of antibody (Supplementary Table 1) and 60 μl of Protein G magnetic beads (Invitrogen). The antibodies were pre-coupled to the beads for 1 h at room temperature with mixing before incubating the precleared lysate to the antibody-bead complexes for 2 h at 4 °C. After the immunoprecipitation, the beads were rinsed with a wash buffer of 1 × PBS with detergents. After a dephosphorylation treatment, the RNA in each sample was ligated to a mixture of barcoded adapters in which each adaptor had a unique barcode identifier. After ligation, beads were rinsed with 1 × PBS and detergents and then 5 × PBS (750 mM NaCl) and detergents before pooling 3–4 antibodies in a new tube. The proteins and RNA were then eluted from the Protein G beads with 6 M urea and 40 mM DTT at 60 °C. Protein–RNA complexes were separated away from free RNA by covalently coupling proteins to NHS-magnetic beads (Pierce) and washing 3 times in 6 M GuSCN (Qiagen RLT buffer) and heating in 1% NLS at 98 °C for 10 min. The proteins were then digested with Proteinase K and RNA was purified for subsequent analysis. From the barcoded RNA in each pool, we generated Illumina sequencing libraries as previously described[34]. We saved a small percentage (~1%) of starting material before immunoprecipitation and processed and sequenced this sample in parallel.

Analysis of crosslinked RNA immunoprecipitation data. We computed the enrichment for any RNA upon immunoprecipitation with a specific protein relative to its total levels in the cell. To do this, we counted the total number of reads overlapping the RNA in either the immunoprecipitation (IP) sample or the input control. To account for differences in read coverage between samples, each of these numbers was normalized to the total number of reads within the same experiment. This generates a normalized score, per RNA, within each sample. We then computed an enrichment metric by taking the ratio of these normalized values (IP/input). We then compared these enrichment levels across different proteins and controls (that is, IgG). To enable direct comparison across proteins for a given gene, we need to account for differences in the protein specific background level, which may occur to differences in IP efficiency or non-specific binding of each antibody. To do this, we computed a normalized enrichment ratio by dividing the ratio for each gene by the average ratio across all genes for a given protein, as previously described[1].

To exclude the possibility of promiscuous binding to all RNAs, we considered various mRNA controls, which are not expected to bind to these proteins, including *Oct4*, *Nanog*, *Stat3*, and *Suz12*. These mRNAs were selected as examples because they are expressed in ES cells, although many mRNAs show similar results. To account for the possibility that the *Xist* RNA non-specifically binds to any RBP, we evaluated *Xist* with other RBPs that we did not identify as interacting with *Xist* by RAP-MS (PUM1 and HNRNPH). To ensure that a negative result (that is, no enrichment for *Xist*) is meaningful and does not reflect a failed immunoprecipitation experiment, we evaluated Neat1-1, which we previously found immunoprecipitates with hnRNPH[1]. To further evaluate the level of enrichment on other lncRNAs, we considered several lncRNAs including Malat1, Firre, and Tug1. These lncRNAs were selected as examples because they are well-known and expressed in ES cells, although many ES lncRNAs show similar results.

Immunoprecipitation and RT-qPCR. Female ES cells were differentiated then crosslinked with UV4k as described above. Pellets of 20 million cells were lysed and treated with TURBO DNase (Ambion) to destroy DNA by incubation for 10 min at 37 °C in an Eppendorf Thermomixer C. The lysate was pre-cleared by incubation with 180 μl of Dynabeads Protein G magnetic beads (Life Technologies). Meanwhile, 10 μg of antibody for immunoprecipitation (SHARP antibody, Novus NBP1-82952 or IgG antibody, Cell Signaling 2729S) was coupled to 60 μl Protein G magnetic beads. After pre-clearing was completed, the lysate was then mixed with the appropriate antibody-coupled Protein G magnetic beads and incubated for 2 h at 4 °C on a Hulamixer sample mixer (Life Technologies) for protein capture. After immunoprecipitation, beads were washed with a wash buffer of 1 × PBS with detergents and then captured nucleic acids were eluted by digesting all proteins with 5.6 U proteinase K (New England Biolabs). Eluted RNA was purified using the RNA Clean and Concentrator-5 Kit (Zima Research) and RT-qPCR was performed as described previously[13] to evaluate RNA enrichment.

V5-epitope tagged protein expression. For V5-tagged protein expression and immunoprecipitation, mouse ES cells were electroporated using the Neon transfection system (Invitrogen) with an episomally-replicating vector (pCAG-GW-V5-Hygro) encoding expression of a C-terminal V5 tagged ORF driven by a CAG promoter. ORFs were obtained from the DNASU plasmid repository as Gateway entry clones and inserted into pCAG-GW-V5-Hygro using an LR recombination reaction (Invitrogen). Transfected cells were selected on 125 μg ml^{-1} Hygromycin B (Invitrogen) to generate stably expressing lines.

siRNA transfections. For siRNA knockdown experiments, 20 nM siRNAs were transfected using the Neon transfection system (settings: 1,200 V, 40 ms width, 1 pulse). For each transfection, two 10 μl transfections with the same siRNA were carried out in succession using 100,000 cells each, mixed, and plated equally between two poly-L-lysine or poly-D-lysine (Sigma) and 0.2% gelatin (Sigma)-coated no. 1.5 coverslips placed into wells of a 24-well plate containing 2i media. After 48 h, 2i media was replaced and cells on one coverslip of each pair were treated with 2 μg ml^{-1} doxycycline (Sigma) for 16hr to induce *Xist* expression. Coverslips were then fixed in Histochoice (Sigma) for 5 min, washed thoroughly in PBS, and dehydrated in ethanol for storage until FISH staining.

For all proteins we used siRNA pools from Dharmacon (ON-TARGETplus SMARTpool siRNAs). For each of these, we tested whether the siRNA successfully reduced the targeted mRNA expression by >70%. For SAF-A and SMRT, the siRNAs failed to achieve this level of mRNA reduction, so we purchased additional siRNAs (and their associated controls) for SAF-A and SMRT from Qiagen and Ambion respectively, and selected siRNAs that successfully reduced on-target mRNA levels. siRNA against GFP was purchased from Qiagen. For additional independent siRNAs, the siRNAs were purchased as a pool from Dharmacon, Qiagen, and Ambion, or as each individual siRNA deconvoluted from the pool from Dharmacon and Qiagen (Supplementary Table 2).

In addition to the proteins identified by RAP-MS, we knocked down several proteins previously reported to associate with *Xist*, including EED (a component of PRC2)[8], YY1[41], SATB1[42], SRSF1[43], HNRNPC[20], and ATRX[21].

siRNA experiments in female ES cells. Female ES F1 2-1 cells were similarly transfected. To initiate differentiation and *Xist* expression for these cells, 2i was replaced with MEF media (DMEM, 10% Gemini Benchmark FBS, 1× L-glutamine, 1× NEAA, 1× penicillin/streptomycin; Life Technologies unless otherwise indicated) at 12 h post-transfection. Forty-eight hours after transfection, 1 μM retinoic acid (Sigma) was administered for 24 h and cells were fixed as described above. For cells not undergoing differentiation, 2i was replaced 12 h and 48 h after transfection.

Single-molecule RNA FISH. Single-molecule RNA fluorescence *in situ* hybridization (FISH) experiments were done using QuantiGene ViewRNA ISH Cell Assay (Affymetrix) and QuantiGene ViewRNA ISH Cell 740 Module (Affymetrix) according to manufacturer's protocol. Cells fixed on coverslips were first permeabilized with Detergent Solution QC at room temperature for 5 min, and then incubated with desired mixture of probe set (Affymetrix) in Probe Set Diluent QF at 40 °C for 3 h, followed by incubated with PreAmplifier Mix at 40 °C for 30 min, Amplifier Mix at 40 °C for 30 min, and Label Probe Mix at 40 °C for 30 min sequentially. For DAPI staining, coverslips were incubated in 30 nM DAPI in PBS at room temperature for 15–20 min. Probe set and conjugated fluorophore for FISH were TYPE 1-XIST (550 nm), TYPE 4-GPC4, RBMX, SMC1A, MECP2 (488 nm), TYPE 10-ATRX (740 nm), and TYPE 6-EED1, SHARP, LBR, SAFA, RBM15, MYEF2, PTBP1, HNRNPC, HNRNPM, CELF1, RALY, HDAC3, NCOR2, MID1, PIR (650 nm).

Immunofluorescence and RNA FISH. For immunofluorescence (IF), cells were fixed on coverslips and permeabilized with 0.1% Triton-X in PBS at room temperature for 10 min, and blocked with 5% normal goat serum in PBS at room temperature for 10 min. Cells were then incubated with primary antibodies at room temperature for 1 h, followed by incubating with secondary antibodies at room temperature for 1 h. The samples were then processed using the RNA FISH protocol, as described above. Primary antibodies and the dilution used for IF were anti-RNA polymerase II CTD repeat YSPTSPS (phospho S2) (Abcam; ab5095) (1:100), anti-NANOG (Abcam; ab80892) (1:100), and anti-EZH2 (Active Motif; 39933) (1:100). Secondary antibodies and the dilution used for IF were Alexa Fluor 405 goat anti-rabbit IgG (H+L) (Life Technology; 1575534) (1:100) and Alexa Fluor 488 $F(ab')_2$ fragment of goat anti-rabbit IgG (H+L) (Life Technology; 1618692) (1:100).

Microscopic imaging. FISH and IF/FISH samples were imaged using a Leica DMI 6000 Deconvolution Microscope with the Leica HC PL APO ×63/1.30 GLYC CORR CS2 objective. Samples stained with TYPE 10-ATRX (740 nm) were

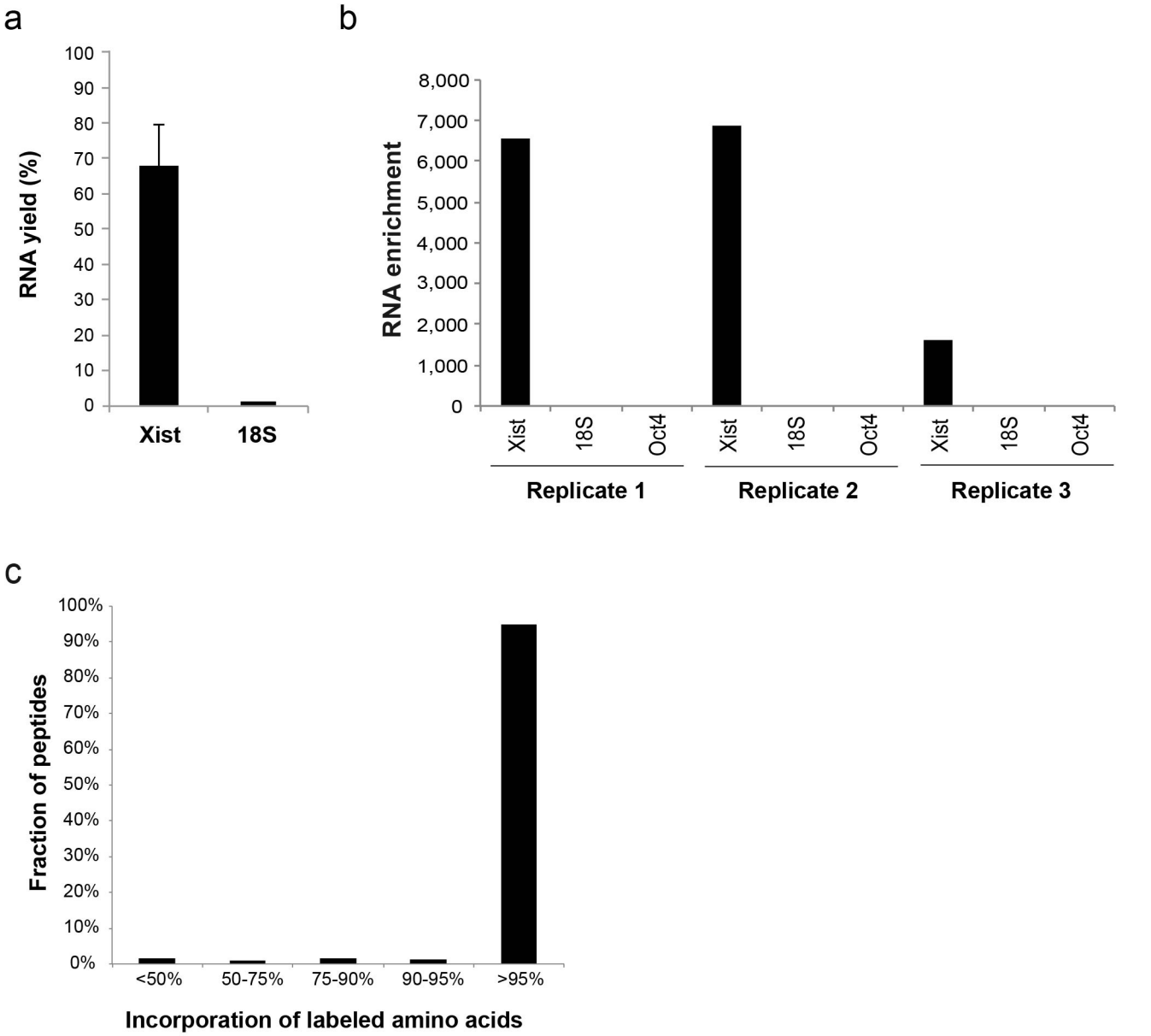
imaged using Nikon Ti Eclipse with the Nikon CFI Plan Apochromat λ DM ×60/1.40 oil objective. Images were projected with maximum projection (3 μm; step size, 0.5 μm).

X-chromosome silencing assay. Cells were stained for *Xist* RNA, *Gpc4* mRNA, *Atrx* mRNA and siRNA-targeted mRNA by FISH and imaged. In addition, in some siScramble and siSHARP samples, we used probes against *Rbmx*, *Mecp2*, *Smc1a*, *Mid1* or *Pir* mRNA. Images were then analysed using Matlab R2013b (described below). Cells were selected if the copy number of the targeted mRNA was less than 30% of the level of the no siRNA treated cells and if they induced *Xist* expression. Within these cells, the copy number of *Gpc4* mRNA and *Atrx* mRNA were quantified using a peak finding method (described below) and compared across conditions. We quantified mRNA levels for 50 individual cells. We also evaluated *Xist* expression in siRNA-treated cells, and observed no difference in the percentage of cells that induced *Xist* expression in any of the siRNA conditions relative to untreated cells.

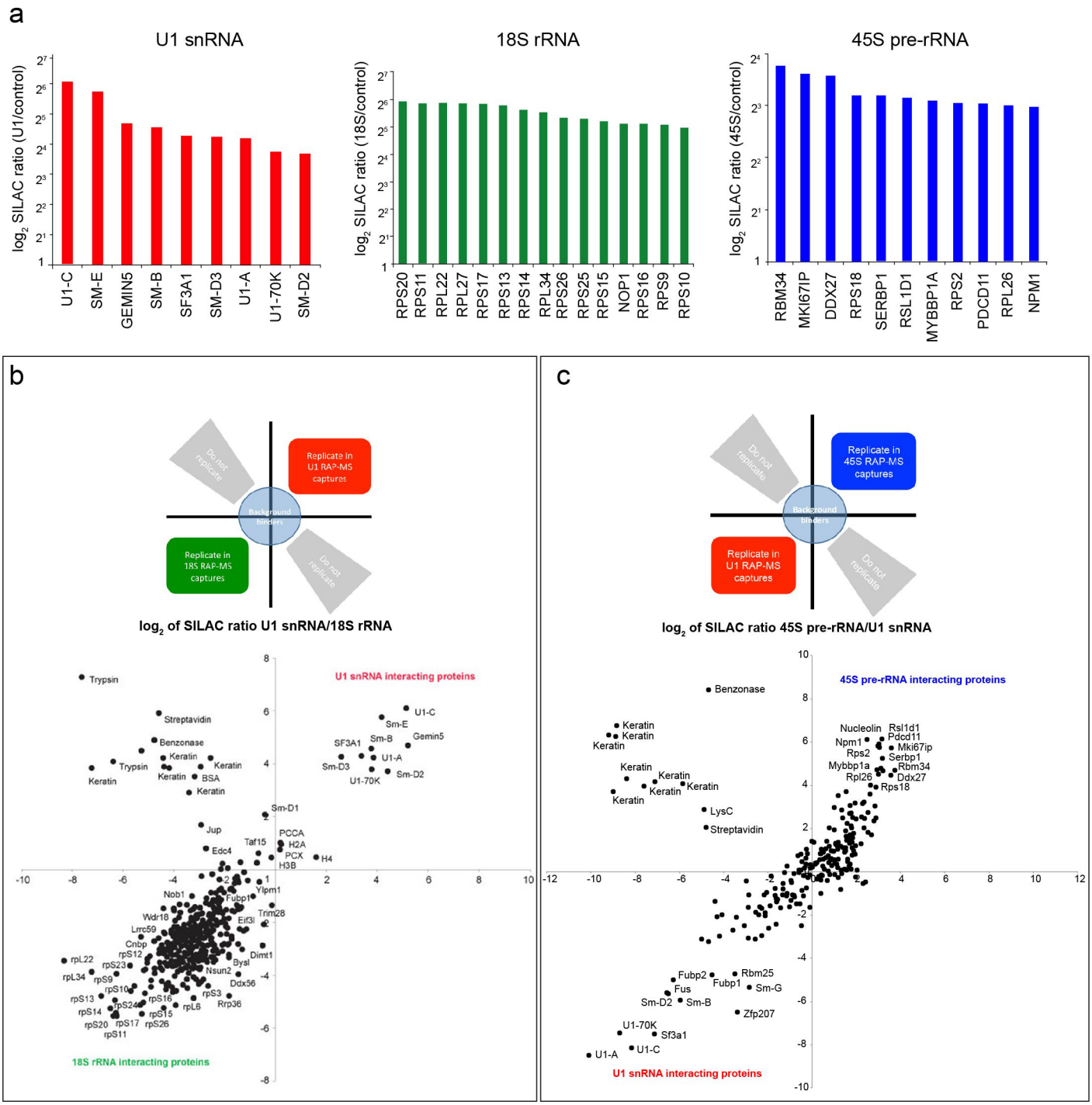
Quantifying mRNAs by single molecule FISH. All image analysis was carried out using Matlab (version R2013b) using built-in functions from the Image Processing toolbox. Images were first filtered using a two-dimensional median filter to remove background. Cell boundaries were outlined manually, guided by DAPI staining, to create a binary mask and applied to the various channels from the same field of view. Top-hat morphological filtering, a background subtraction method that enhances the individual focal spots, was applied to the images[44]. The spots were then identified using a 2D peak finding algorithm that identifies local maximal signals within the cell. Once regional maxima were identified, the number of spots was counted for each cell.

Ezh2 recruitment and Pol II exclusion. Cells were stained for *Xist* RNA and the siRNA-targeted mRNA (FISH) along with EZH2 or Pol II (IF) as described above. For image acquisition, the exposure time for each individual channel was kept the same across all samples. Images were then analysed and selected for XIST-induced and cells showing knockdown of the target mRNA, as described above. Specifically, the nuclei of individual cells were identified manually using the DAPI staining. We identified the *Xist*-coated territory by using an intensity-based threshold to partition the image within the nucleus and find contiguous 2-dimensional regions of high intensity. The threshold was determined based on the Otsu method as previously described[45], which splits the image into 2 bins—high and low—and identifies a threshold that minimizes the variance within the partition. This creates a binary mask on the image. We visually confirmed that this binary mask accurately reflected the *Xist*-coated territory. We then applied this binary mask to all other images in that field of view (Pol II or EZH2) for all images. We then quantified the intensity of fluorescence signal by taking the average intensity of all the pixels within the region (that is, Pol II or EZH2, respectively). We computed this average intensity (1 number per cell) across all conditions and compared them using a 2-same unpaired *t*-test relative to the scramble sample across 50 single cells.

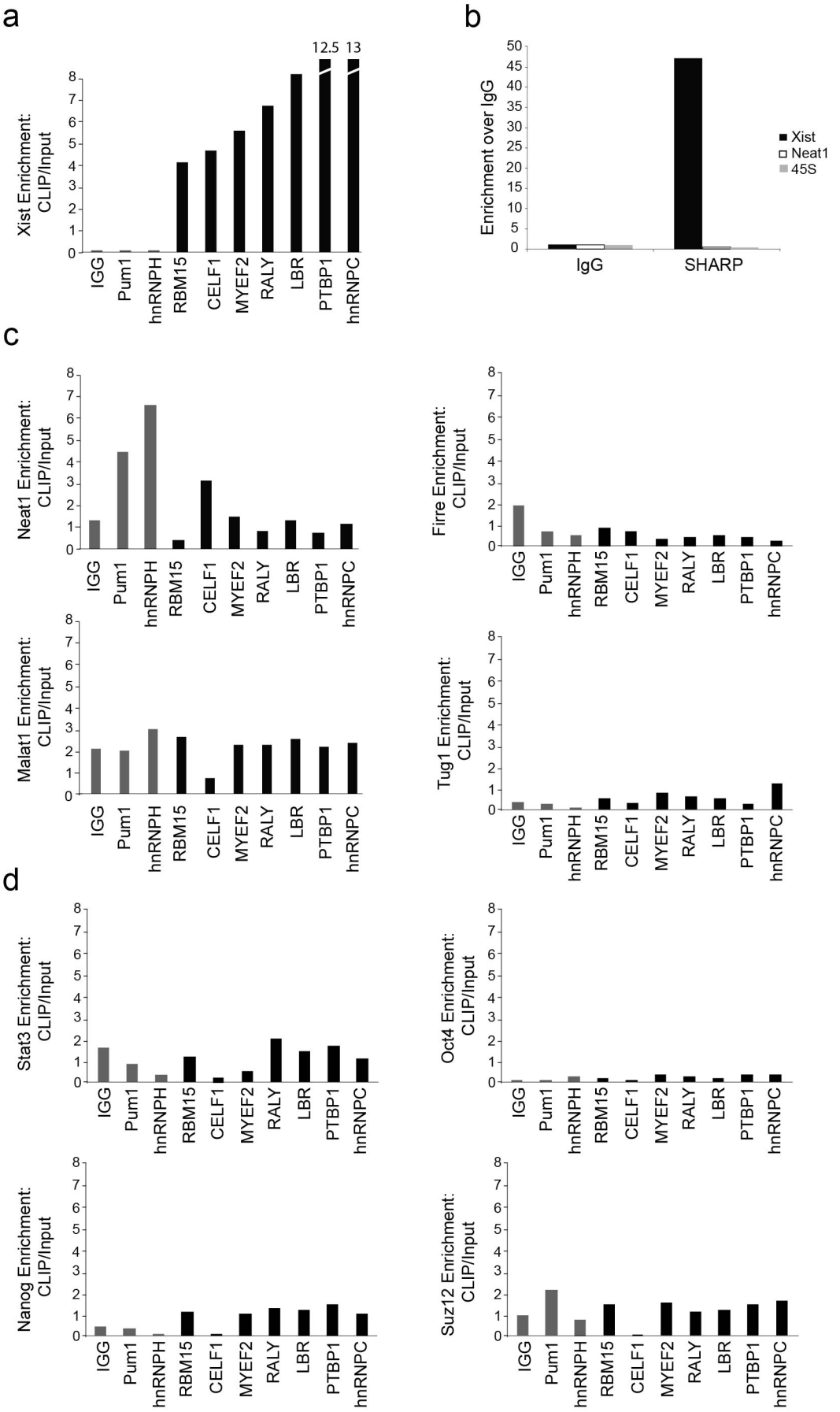
33. Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nature Genet.* **30,** 167–174 (2002).
34. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159,** 188–199 (2014).
35. Kalli, A. & Hess, S. Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *Proteomics* **12,** 21–31 (2012).
36. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372 (2008).
37. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).
38. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **604,** 55–71 (2010).
39. Geiger, T. *et al.* Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature Protoc.* **6,** 147–157 (2011).
40. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
41. Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146,** 119–133 (2011).
42. Agrelo, R. *et al. SATB1* defines the developmental context for gene silencing by *Xist* in lymphoma and embryonic cells. *Dev. Cell* **16,** 507–516 (2009).
43. Royce-Tolland, M. E. *et al.* The A-repeat links ASF/SF2-dependent Xist RNA processing with random choice during X inactivation. *Nature Struct. Mol. Biol.* **17,** 948–954 (2010).
44. Theodosiou, Z. *et al.* Automated analysis of FISH and immunohistochemistry images: a review. *Cytometry A* **71,** 439–450 (2007).
45. Fumagalli, M. *et al.* Telomeric DNA damage is irreparable and causes persistent DNA-damage-response activation. *Nature Cell Biol.* **14,** 355–365 (2012).



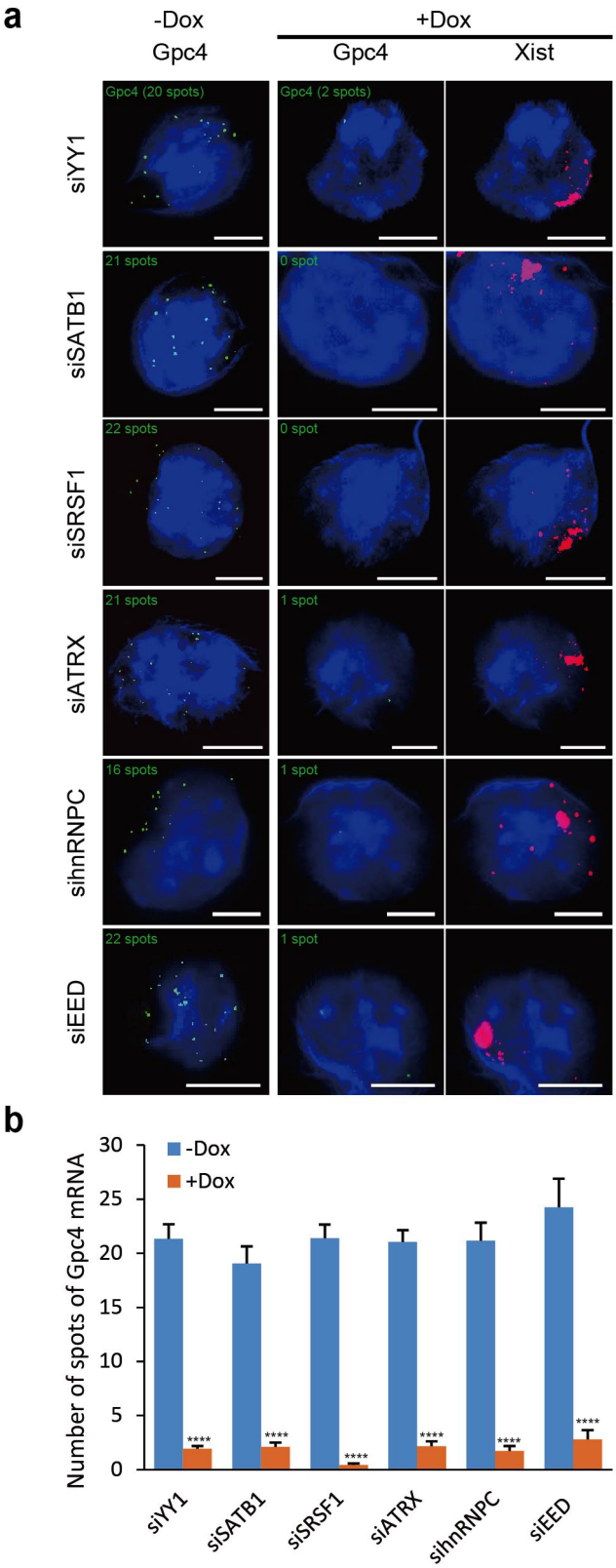
Extended Data Figure 1 | RAP-MS recovers and enriches the majority of *Xist* RNA from mouse ES cells, and these cells can be efficiently labelled with SILAC. **a**, RT-qPCR measuring the percentage of the total cellular *Xist* or 18S recovered after RAP-MS of *Xist*. Values are computed as the amount of each RNA in the elution divided by the amount of RNA in the starting ('input') lysate material. Error bars represent the standard error of the mean from 5 biological replicates. **b**, Enrichment of *Xist* after RAP-MS captures from pSM33 cells as measured by qPCR. Bars indicate RNA levels of *Xist*, 18S, and Oct4 after purification of *Xist*, normalized to RNA in input sample. Each bar represents the RNA levels of *Xist*, 18S, and Oct4 after purification of *Xist*, normalized to RNA in input sample, from 3 biological replicates. **c**, SILAC labelling efficiency of a representative culture of pSM33 mouse ES cells after 10 days of growth (3 cell passages) in SILAC medium. Peptides were analysed by mass spectrometry, and values indicate the fraction of identified peptides with heavy-label incorporation with different levels of peptide labelling (shown in bins).



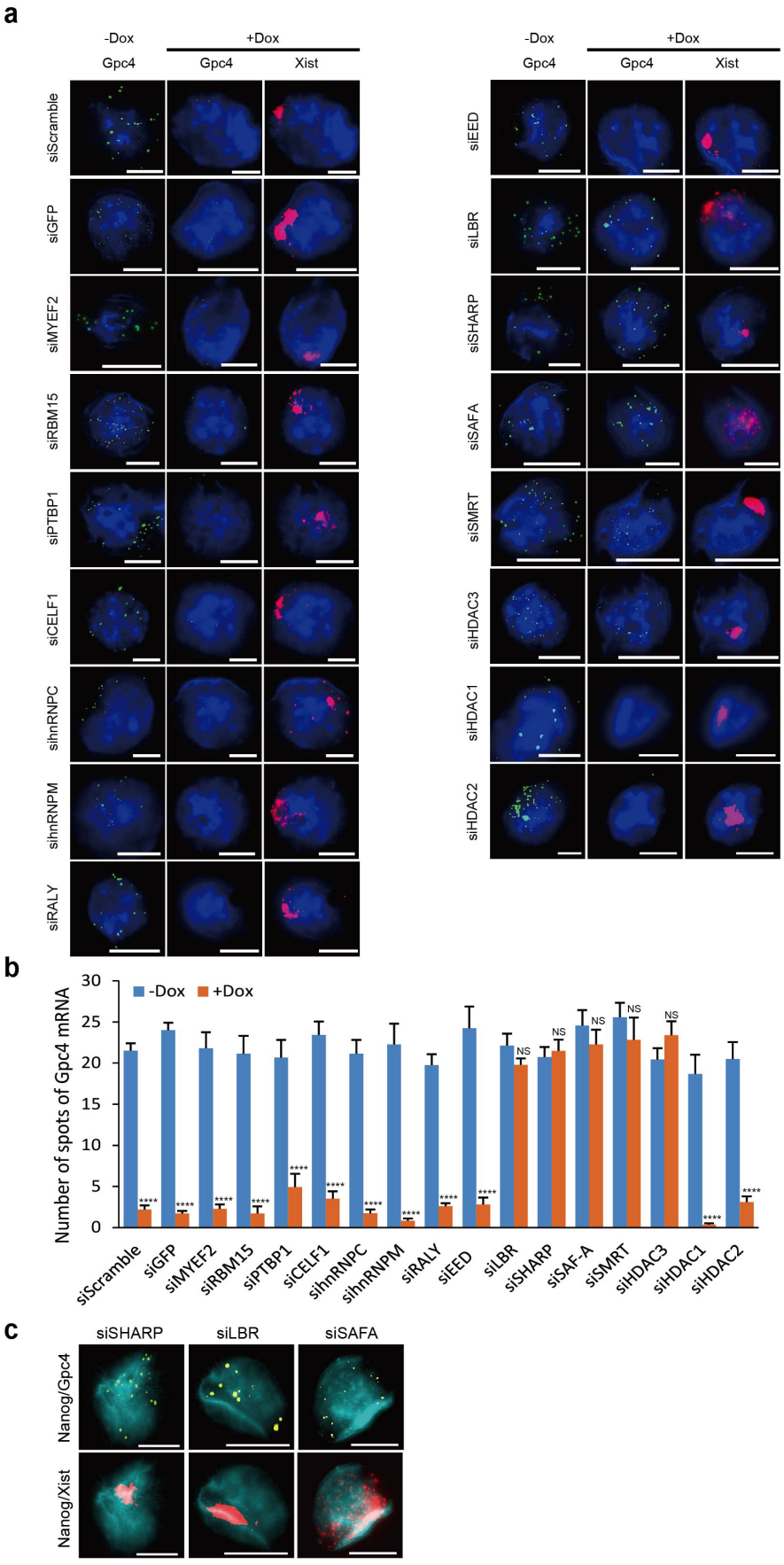
Extended Data Figure 2 | RAP-MS identifies proteins that are known to directly interact with specific ncRNAs, and separates specific RNA interacting proteins from background proteins. **a**, SILAC ratios of top proteins enriched in the RAP-MS U1 snRNA, 18S rRNA, and 45S pre-rRNA experiments. **b**, SILAC ratio plot of replicate captures of U1 snRNA versus 18S rRNA from one of two biologically independent label-swap experiments. Proteins associated with U1 are consistently found in U1 samples, both light and heavy labelled (top right quadrant), and proteins specifically associated with 18S are consistently identified in 18S, both light and heavy (lower left quadrant). Background contaminant proteins have low enrichments (centre of panel) or are consistently found in the light channel and do not replicate between experiments (that is, keratin, streptavidin). **c**, SILAC ratio plot of replicate captures of U1 snRNA versus 45S pre-rRNA from one label-swap experiment. Proteins that are known to associate with 45S pre-rRNA are consistently identified in 45S captures.



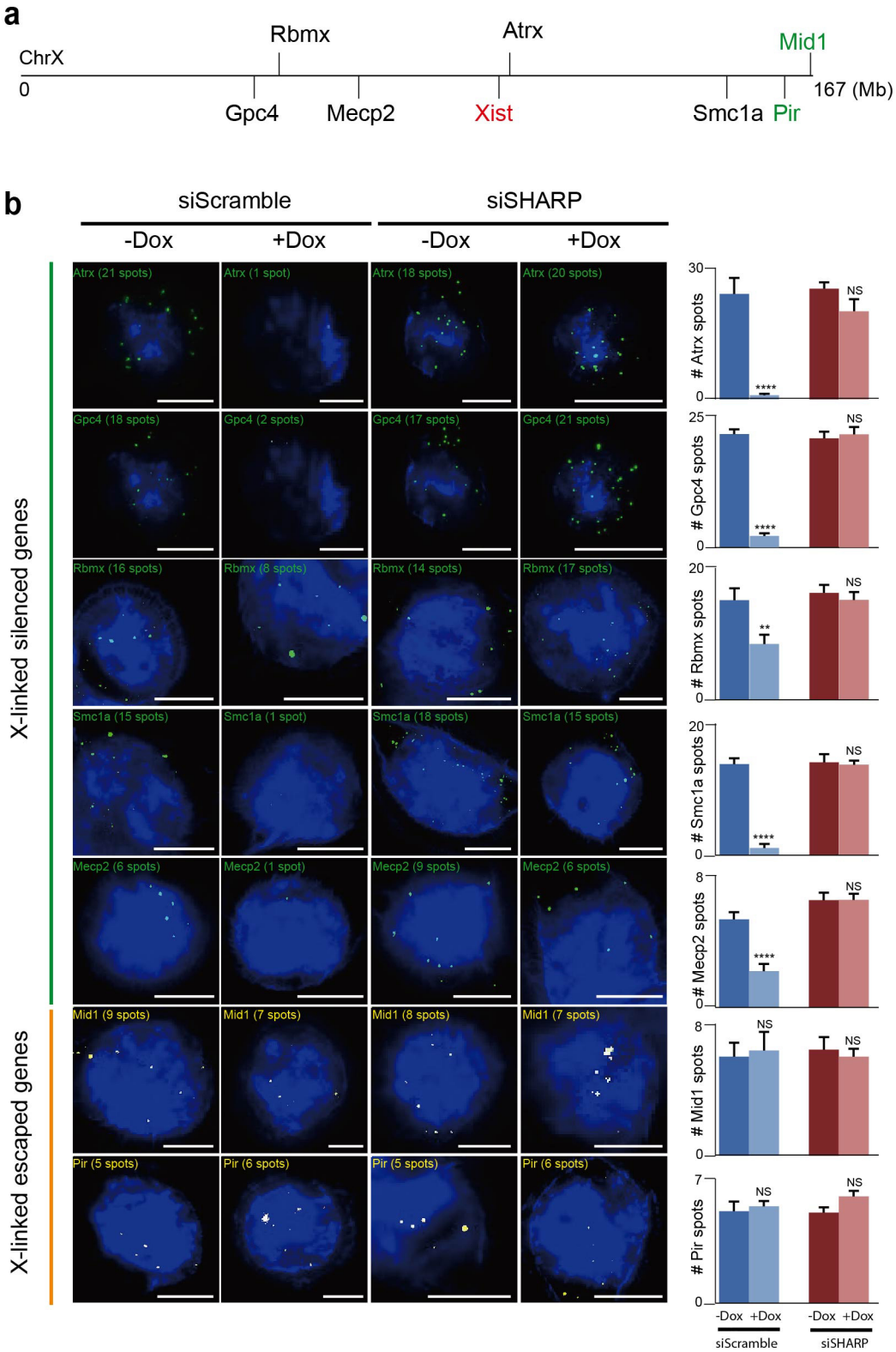
Extended Data Figure 3 | Immunoprecipitation of the identified *Xist*-interacting proteins confirms *Xist* RNA interaction. RNA immunoprecipitation experiments were performed for seven *Xist*-interacting proteins (black bars), two control RNA binding proteins that were not identified by RAP-MS and IgG (grey bars) in UV-crosslinked cell lysate after 6 h of *Xist* induction by doxycycline addition (Methods). The RNA associated with each protein was measured and enrichment levels were computed relative to the level of the RNA in total cellular input and normalized to the total efficiency of capture in each sample to allow for direct comparison across all immunoprecipitation experiments (Methods). **a**, Enrichment of the *Xist* lncRNA after immunoprecipitation from a sample of pSM33 male cells. **b**, Immunoprecipitation of SHARP was performed from a sample of UV-crosslinked females ES cells that were treated with retinoic acid for 24 h. The levels of recovered *Xist* lncRNA (black bars), Neat1 lncRNA (white bars), and 45S pre-ribosomal RNA (grey bars) were measured by RT-qPCR. Enrichment of each RNA after capture with anti-SHARP antibody was calculated relative to the level of RNA captured with IgG control antibody. **c**, The enrichment of various lncRNAs after immunoprecipitation in pSM33 male cells—including Neat1, Malat1, Firre, and Tug1—are shown. **d**, The enrichment of various mRNA controls after immunoprecipitation in pSM33 male cells—including Oct4, Nanog, Stat3, and Suz12—are shown.



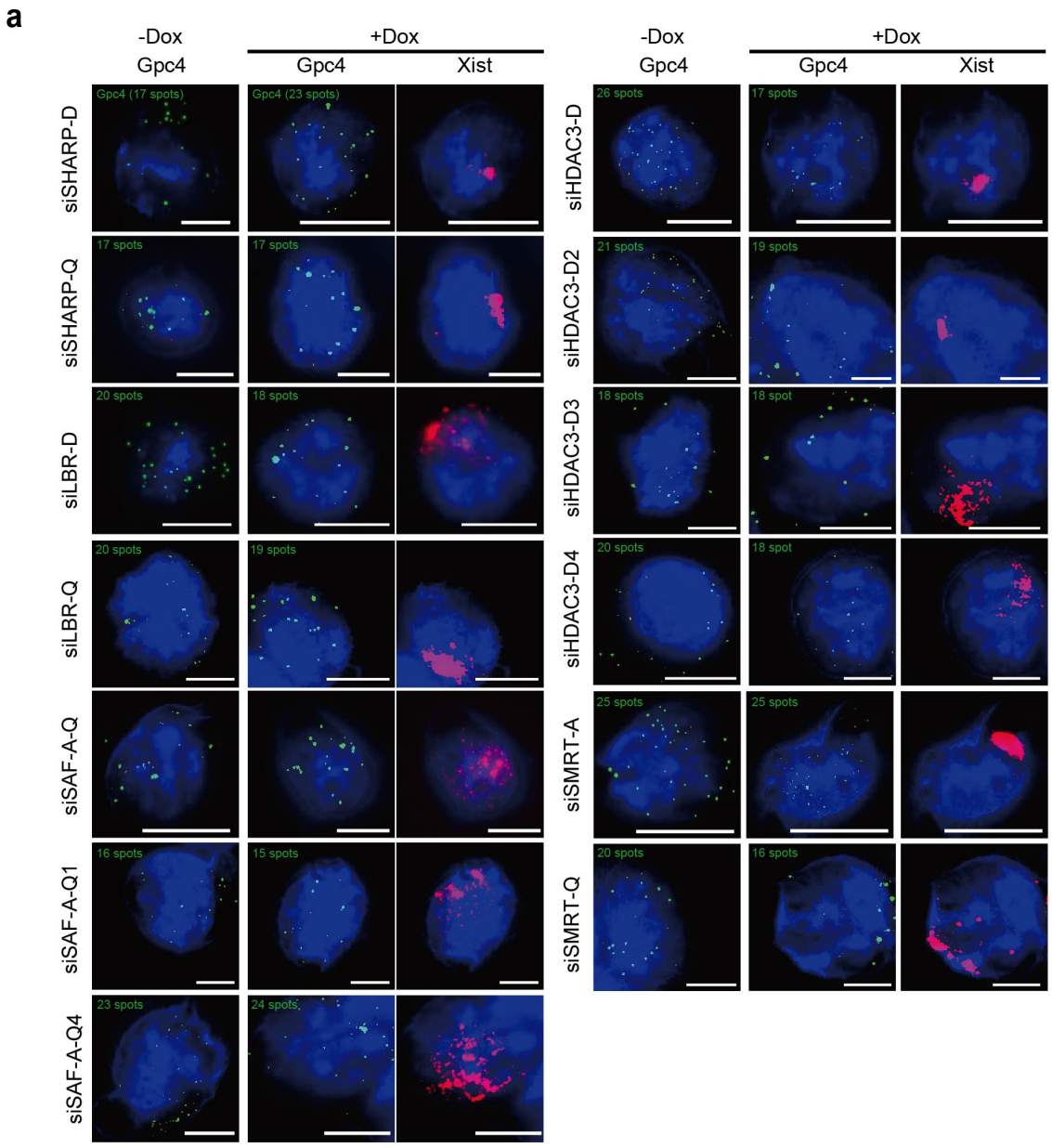
Extended Data Figure 4 | Previously identified proteins associated with XCI are not required for *Xist*-mediated transcriptional silencing. **a**, To confirm the specificity of our assay, we tested the function of several proteins that were previously identified to associate with *Xist*, but not to silence transcription, for their role in transcriptional silencing in our inducible male ES cells before *Xist* induction (−Dox; left) or after *Xist* induction for 16 h (+Dox; middle and right). Representative images are shown after knockdown of each protein. DAPI (blue), *Xist* (red), and Gpc4 (green). **b**, Quantification of the copy number of Gpc4 before and after *Xist* induction upon treatment with different siRNAs. Error bars represent the standard error of the mean across 50 individual cells from one experiment. ****P value < 0.001 between +Dox and –Dox cells based on an unpaired two-sample t-test. Scale bars on the images represent 5 μm. Importantly, while these proteins do not have a role in the initiation of transcriptional silencing, we do not mean to imply that they do not have other roles in XCI.



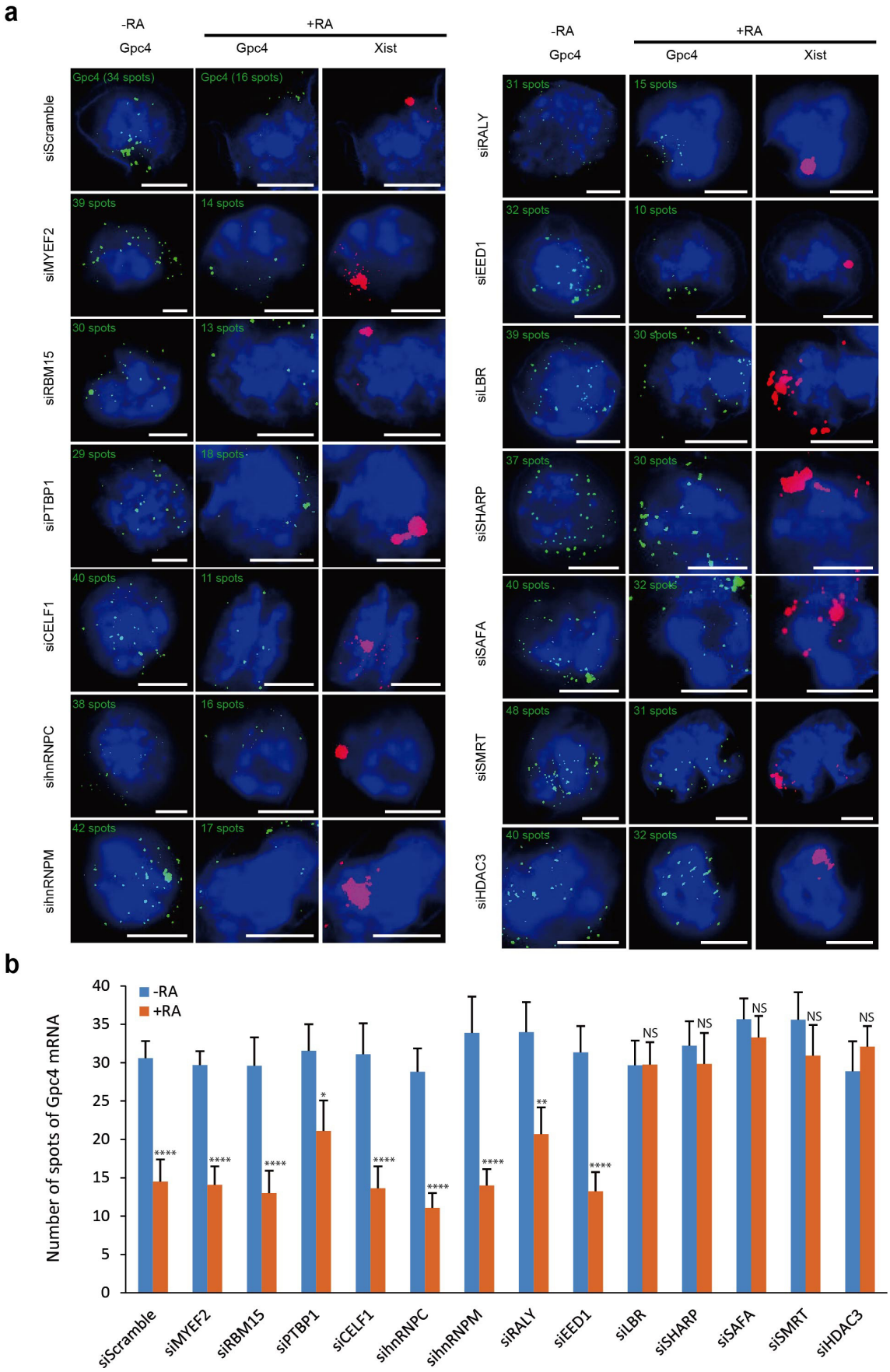
Extended Data Figure 5 | SHARP, LBR, SAF-A, SMRT, and HDAC3 are required for *Xist*-mediated transcriptional silencing. **a**, Representative images showing staining of DAPI (blue), *Xist* (red), and Gpc4 (green) for different siRNA knockdown in male ES cells before *Xist* induction (−Dox; left) or after *Xist* induction for 16 h (+Dox; middle and right). **b**, Quantification of the copy number of Gpc4 in –Dox and +Dox cells after knockdown with siRNAs targeting different mRNAs. Error bars represent the standard error of the mean across 50 individual cells from one experiment. NS, not significantly different between +Dox and –Dox cells; ****P value < 0.001 between +Dox and –Dox cells based on an unpaired two-sample t-test. Scale bars on the images represent 5 μm. **c**, Knockdown of SHARP, LBR, or SAF-A abrogates *Xist*-mediated gene silencing without causing pluripotency defects. Representative images showing staining of Nanog (cyan), *Xist* (red), and Gpc4 (green) upon knockdown of SHARP, LBR or SAF-A after 16 h of *Xist* induction with doxycycline. Scale bars on the images represent 5 μm.



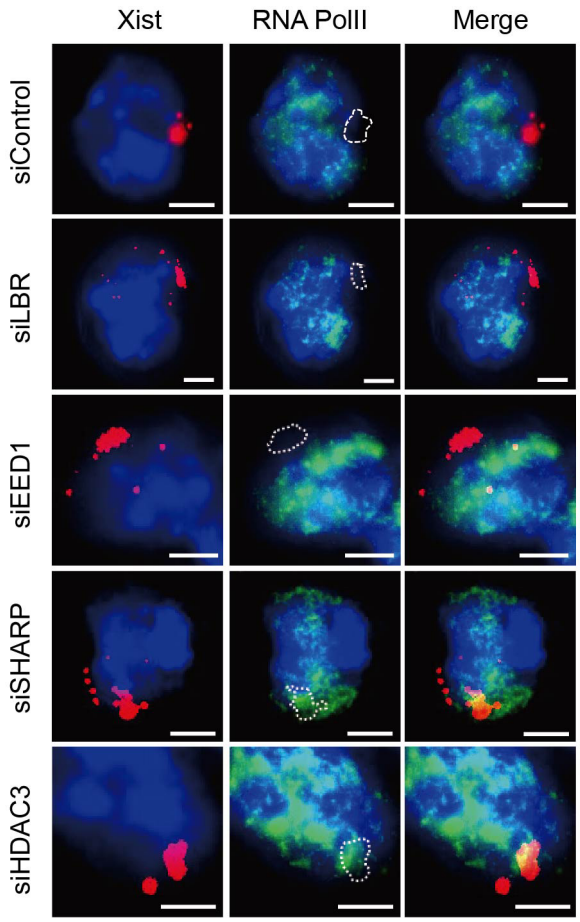
Extended Data Figure 6 | SHARP is required for silencing many genes across the X chromosome. **a**, A diagram showing the locations of *Xist* (red), X-linked silenced genes (black), and X-linked escaped genes (green) along the X chromosome. **b**, Representative images showing staining of DAPI (blue), *Xist* (red), X-linked silenced genes (green), and X-linked escaped genes (yellow) upon knockdown of SHARP or control male ES cells before *Xist* induction (−Dox) or after *Xist* induction for 16 h (+Dox). Knock of SHARP abolishes the silencing of Atrx, Gpc4, Rbmx, Smc1a and Mecp2, which are normally silenced upon *Xist* expression, but has no effect on Mid1 and Pir, which normally escape *Xist*-mediated silencing. The bar graphs show the quantification of the copy number of the mRNA for each gene for –Dox and +Dox cells upon transfection with SHARP siRNA or control siRNA; error bars represent the standard error of the mean across 50 individual cells from one experiment. NS, not significantly different, ****P value < 0.001, and **P value < 0.01 between +Dox and –Dox cells based on an unpaired two-sample t-test. Scale bars on the images represent 5 μm.



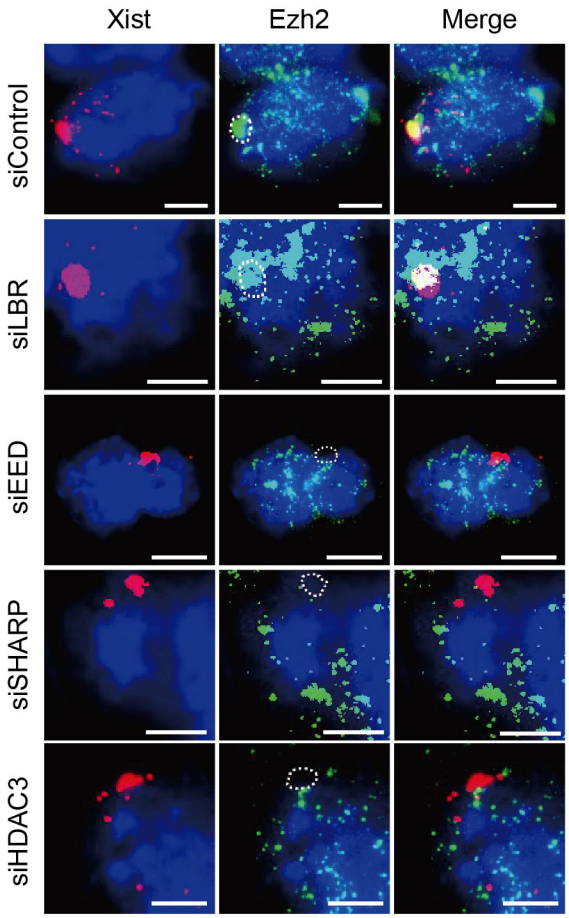
Extended Data Figure 7 | Multiple independent siRNAs targeting SHARP, LBR, SAF-A, HDAC3, or SMRT demonstrate the same silencing defect. **a**, Representative images showing staining of DAPI (blue), *Xist* (red), and *Gpc4* (green) after knockdown of proteins using independent, non-overlapping, siRNA pools, or individual siRNA deconvoluted from the pool before *Xist* induction (−Dox; left) or after *Xist* induction for 16 h (+Dox; middle and right). Cells were either transfected with the siRNA pool from Dharmacon (siRNA-D), Qiagen (siRNA-Q) or Ambion/Life Technologies (siRNA-A), or each individual siRNA deconvoluted from the pool from Dharmacon (siRNA-D1, 2, 3, 4) or Qiagen (siRNA-Q1, 2, 3, 4). **b**, Quantification of the copy number of *Gpc4* in –Dox and +Dox cells after knockdown with siRNAs targeting different mRNAs. Error bars represent the standard error of the mean across 50 individual cells from one experiment. NS, not significantly different between +Dox and –Dox cells based on an unpaired two-sample *t*-test. Scale bars on the images represent 5 μm. We excluded all siRNAs that did not reduce the targeted mRNA level by >70% (Methods). The sequences of deconvoluted siRNAs are shown in Supplementary Table 2.



Extended Data Figure 8 | SHARP, LBR, SAF-A, SMRT, and HDAC3 are required for transcriptional silencing in differentiating female ES cells. **a**, Representative images showing staining of DAPI (blue), *Xist* (red), and Gpc4 (green) upon knockdown of specific proteins using different siRNAs in female ES cells before differentiation (−RA; left) or after differentiation for 24 h (+RA; middle and right). RA, retinoic acid. **b**, Quantification of the copy number of Gpc4 for –RA and +RA cells upon transfection with different siRNAs. Error bars represent the standard error of the mean across 50 individual cells from one experiment. NS, not significantly different between +RA and –RA cells; ****P value < 0.001, **P value < 0.01, and *P value < 0.05 between +RA and –RA cells based on an unpaired two-sample t-test. Scale bars on the images represent 5 μm.



Extended Data Figure 9 | SHARP is required for exclusion of RNA polymerase II from the *Xist*-coated territory in differentiating female ES cells. Images of individual cells that are labelled with *Xist* (red), RNA Polymerase II (green), and DAPI (blue) across different siRNA conditions (rows) in female ES cells after 24 h of retinoic acid treatment. The dashed white region represents the outlined *Xist*-coated territory.



Extended Data Figure 10 | SHARP is required for PRC2 recruitment across the *Xist*-coated territory in differentiating female ES cells. Images of individual cells that are labelled with *Xist* (red), Ezh2 (green) and DAPI (blue) across different siRNA conditions (rows) in female ES cells after 24 h of differentiation. The dashed white region represents the outlined *Xist*-coated territory.

LETTER

doi:10.1038/nature14185

Horizontal membrane-intrinsic α-helices in the stator *a*-subunit of an F-type ATP synthase

Matteo Allegretti[1], Niklas Klusch[1], Deryck J. Mills[1], Janet Vonck[1], Werner Kühlbrandt[1] & Karen M. Davies[1]

ATP, the universal energy currency of cells, is produced by F-type ATP synthases, which are ancient, membrane-bound nanomachines. F-type ATP synthases use the energy of a transmembrane electrochemical gradient to generate ATP by rotary catalysis. Protons moving across the membrane drive a rotor ring composed of 8–15 *c*-subunits[1]. A central stalk transmits the rotation of the *c*-ring to the catalytic F_1 head, where a series of conformational changes results in ATP synthesis[2]. A key unresolved question in this fundamental process is how protons pass through the membrane to drive ATP production. Mitochondrial ATP synthases form V-shaped homodimers in cristae membranes[3]. Here we report the structure of a native and active mitochondrial ATP synthase dimer, determined by single-particle electron cryomicroscopy at 6.2 Å resolution. Our structure shows four long, horizontal membrane-intrinsic α-helices in the *a*-subunit, arranged in two hairpins at an angle of approximately 70° relative to the *c*-ring helices. It has been proposed that a strictly conserved membrane-embedded arginine in the *a*-subunit couples proton translocation to *c*-ring rotation[4]. A fit of the conserved carboxy-terminal *a*-subunit sequence places the conserved arginine next to a proton-binding *c*-subunit glutamate. The map shows a slanting solvent-accessible channel that extends from the mitochondrial matrix to the conserved arginine. Another hydrophilic cavity on the lumenal membrane surface defines a direct route for the protons to an essential histidine–glutamate pair[5]. Our results provide unique new insights into the structure and function of rotary ATP synthases and explain how ATP production is coupled to proton translocation.

To investigate the structural basis of proton translocation in F-type ATP synthases, we performed single-particle electron cryomicroscopy on detergent-solubilized mitochondrial ATP synthase dimers of the colourless green alga *Polytomella* sp. Dimers of *Polytomella* ATP synthase have a molecular mass of approximately 1.6 MDa and are particularly stable owing to their bulky peripheral stalk formed by the ATP synthase-associated subunits, ASA1–9 (ref. 6). The complex used for electron cryomicroscopy data collection had high oligomycin-sensitive ATPase activity (Extended Data Fig. 1). A total of 37,238 dimer images were combined to generate a map with an overall resolution of 7.0 Å (Fig. 1, Extended Data Fig. 2 and Supplementary Video 1). Subvolume masking indicated a resolution of 6.2 Å for the rigid peripheral stalk and its associated membrane domains (Extended Data Fig. 3), whereas the resolution for the more flexible catalytic F_1 subcomplex and *c*-ring was 7.4 Å. α-Helices were clearly resolved in all parts of the symmetrized and unsymmetrized map (Extended Data Fig. 4).

The ATP synthase dimer has a twofold symmetrical V-shape. Atomic models of homologous F-type ATP synthase subcomplexes[7,8] were easily fitted into the map (Fig. 1a). The peripheral stalks form a solid scaffold of multiply-entwined, long α-helices. Cross-sections through the membrane-embedded part of the dimer reveal six transmembrane helices and a rotor ring of ten *c*-subunits, each forming a helix hairpin perpendicular to the membrane plane (Fig. 1b, c). The dimer interface extends from the peripheral stalk region through the hydrophobic membrane core to the lumenal membrane surface. About 80 Å above the membrane surface, a helix–turn–helix motif protrudes from the peripheral stalk to interact with its counterpart in the other protomer (Fig. 1d). On the matrix membrane surface, a structural motif resembling Armadillo repeat proteins[9] bridges the two peripheral stalks (Fig. 1e). In the

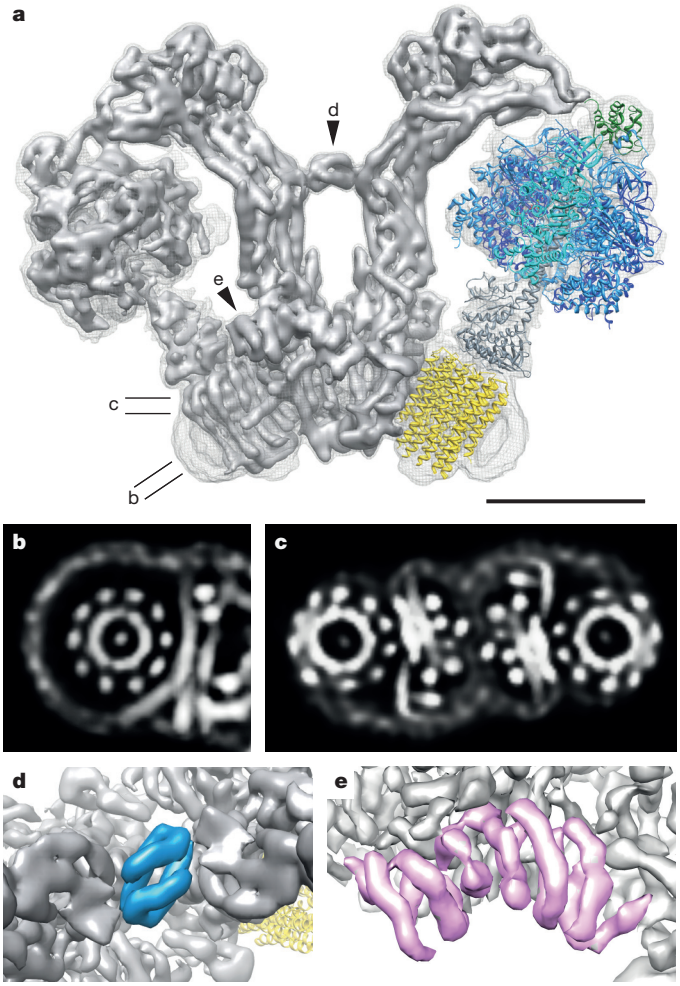


Figure 1 | Electron cryomicroscopy map of the *Polytomella* F-type ATP synthase dimer. **a**, Side view. One half is fitted with atomic models of the F_1 subcomplex (α-subunits, cyan; β-subunits, blue; γδε-subunits, grey; oligomycin-sensitivity-conferring protein, green; Protein Data Bank accession number 2WSS[8]) and the c-ring (yellow, c_{10}-ring; Protein Data Bank accession number 3U2Y[7]). Density threshold levels are 1σ (mesh) or 7σ (solid surface). **b**, **c**, Cross-sections through the membrane domains as indicated in **a**, showing the c_{10} rotor ring surrounded by detergent (diffuse grey density in **b**) and six transmembrane helices per protomer. **c**, Horizontal *a*-subunit helices are visible as elongated white densities. **d**, **e**, Helix–turn–helix motif (blue) and Armadillo repeat-like density (pink) connecting the two protomers, viewed in the direction of the arrowheads in **a**. Scale bar, 100 Å. See also Supplementary Video 1.

[1]Department of Structural Biology, Max Planck Institute of Biophysics, Max-von-Laue-Strasse 3, 60438 Frankfurt am Main, Germany.

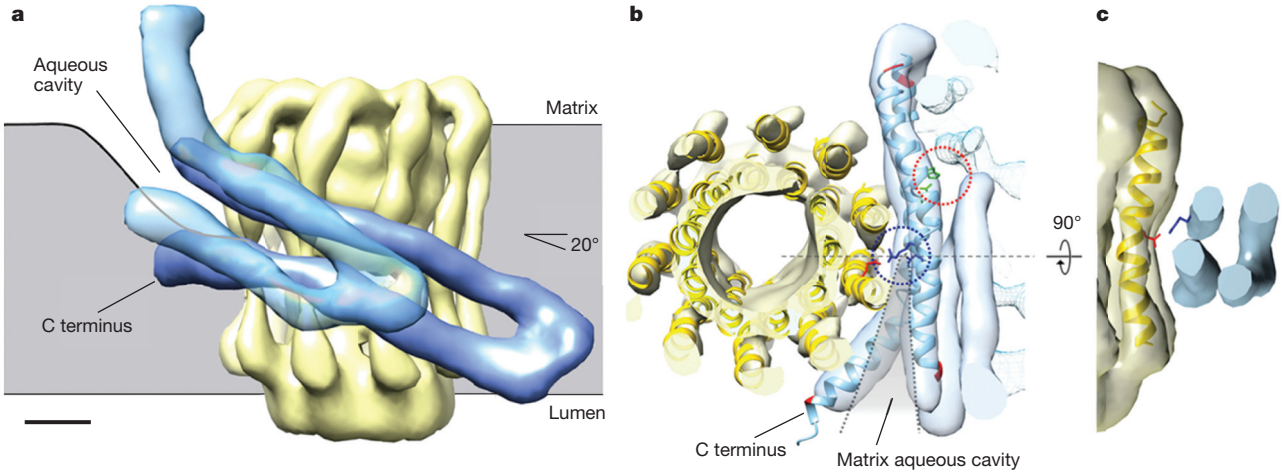


Figure 2 | *a*-Subunit structure. **a**, Membrane-intrinsic helix hairpins of subunit *a* (blue) and c_{10}-rotor ring (yellow) viewed from the dimer interface (see Supplementary Video 2). **b**, Matrix view of the *c*-ring (yellow; fitted atomic model, Protein Data Bank accession number 3U2Y[7]) and C-terminal hairpin of subunit *a* with fitted helices. Blue and red dotted circles indicate the approximate positions of interchangeable residue pairs *a*R239/*a*Q295 (ref. 17) (blue) and *a*E288/*a*H248 (ref. 5) (green). Red ribbon, prolines at helix ends. The dashed grey outline marks the aqueous matrix cavity. **c**, Orthogonal section through *c*-ring and the *a*-subunit four-helix bundle along the dashed line in **b**. In its proposed position, the strictly conserved *a*R239 (dark blue) interacts with proton-binding glutamates of the outer *c*-ring helices (red). Scale bar, 10 Å.

membrane, two transmembrane helices from each protomer line the dimer interface (Fig. 1c). On the lumenal surface, the ends of three transmembrane helices interact to connect the two protomers (Supplementary Video 1).

Between each *c*-ring and the six transmembrane helices, the map reveals a bundle of four long membrane-intrinsic helices arranged in two hairpins, almost at right angles to the *c*-ring (Figs 1b, c and 2 and Supplementary Video 2). The longest (~80 Å) membrane-intrinsic helix is embedded mostly in the lumenal membrane leaflet, where it bends around the *c*-ring for more than one-third of its length. The second helix in this hairpin is ~65 Å long, roughly straight and connected to the curved long helix via a tight loop in the lumenal leaflet (Fig. 2a). Both helices are in contact with the *c*-ring. The other, shorter (~35 Å) membrane-intrinsic hairpin is located between the six transmembrane helices and the longer helix hairpin, and is continuous with a helix density that extends into the peripheral stalk on the matrix side (Fig. 2a). In *Escherichia coli*, the *a*-subunit can be multiply crosslinked to the *c*-ring[10,11]. Since the four horizontal helices are the only protein density next to the *c*-ring in our structure, we assign them to subunit *a*.

The horizontal orientation of the membrane-intrinsic *a*-subunit helices was unexpected. Previous electron cryomicroscopy studies[12,13] and chemical crosslinking[14] had suggested four or more vertical helices in the *a*-subunit. However, our 6.2 Å map of the *Polytomella* dimer is unambiguous in showing that the *a*-subunit helices are oriented horizontally (see Supplementary Discussion and Extended Data Fig. 5).

A comparison of *a*-subunit sequences from bacteria, mitochondria and chloroplasts indicates a conserved carboxy (C)-terminal segment of 90–100 residues (Extended Data Fig. 6), while the remainder is highly divergent. The C-terminal segment includes a strictly conserved arginine (*a*R239) that is essential for coupled proton transfer in *E. coli*[4]. Several residues in this segment can be crosslinked to the *c*-ring in *E. coli*[10,11]. We conclude that the long hairpin contains the conserved arginine and the residues that crosslink to the *c*-ring (Extended Data Fig. 7). We assign the free end of the curved helix (Fig. 2a) to the C terminus of subunit *a*, because the other helix in this hairpin is connected to protein density of the peripheral stalk at a lower contour level.

The two long helices in the C-terminal hairpin comprise about 40 and 50 residues, respectively. Examining the sequence of the *Polytomella* mitochondrial *a*-subunit, we detected two largely hydrophobic stretches of 41 and 47 residues, each with a proline at either end (Extended Data Fig. 6). Proline residues are often found at the start or end of α-helices[15]. We fitted α-helices of the corresponding sequences to the long C-terminal hairpin (Fig. 2b). Similar, proline-delineated stretches are present in all compared *a*-subunit sequences (Extended Data Fig. 6), suggesting that a horizontal helix hairpin next to the *c*-ring is a feature common to all F-type ATP synthases.

Our fit places the essential *a*R239 next to the proton-binding *c*-subunit glutamate[16], at the point of closest proximity between the *c*-ring and *a*-subunit helices (Fig. 2b, c). In *E. coli*, this arginine can be exchanged against glutamine with 20% retention of activity, provided that a conserved glutamine in the second helical stretch is in turn replaced by arginine[17]. In *Polytomella*, the conserved glutamine is *a*Q295. Our fit places this glutamine in the long curved helix directly below the proposed position of *a*R239 (Fig. 2b). In *E. coli*, the equivalents of *a*E248 and *a*H288 form another essential pair of conserved, mutually interchangeable residues[5]. In our fit, *a*E248 and *a*H288 are likewise directly above one another in the long helix hairpin (Fig. 2b). Our model explains why these two residue pairs are interchangeable, and how an arginine in place of *a*Q295 can substitute for *a*R239 in coupled proton translocation.

On the basis of mutagenesis studies in *E. coli*, proton translocation through subunit *a* has been proposed to involve two aqueous half channels[18]. The ends of the two long hairpin helices line a slanting aqueous cavity on the matrix side of the detergent shell. The cavity extends down to the curved hairpin helix and narrows to a slit at the point where the *a*- and *c*-subunits interact in the middle of the membrane (Figs 2a, b and 3a, c and Supplementary Video 3). The aqueous cavity partly exposes the matrix ends of three *c*-subunit helices (Fig. 3a). Protons carried into this cavity by the *c*-subunit glutamates would be rapidly lost to the high pH of the matrix solvent[19]. In our model, the matrix cavity is lined by the amino (N)-terminal part of the long straight helix (residues 225–235), which contains conserved polar or charged residues for three helix turns upstream of *a*R239 (Extended Data Fig. 6). This helix segment is therefore partly hydrophilic, or amphipathic. Cysteine mutation experiments in *E. coli* indicated that residues in this region, as well as residues 44–66 in the *E. coli c*-ring, are solvent-exposed[18,20,21], in excellent agreement with our model (Supplementary Video 4). On the lumenal membrane surface, we find a further cavity that extends from the cristae lumen to the site of the essential *a*H248/*a*E296 pair of interchangeable residues (Fig. 3b, d and Supplementary Video 3). The corresponding residues in *E. coli* have been shown to be ion-accessible[18]. We conclude that this lumenal cavity is the entrance channel for protons crossing the inner mitochondrial membrane.

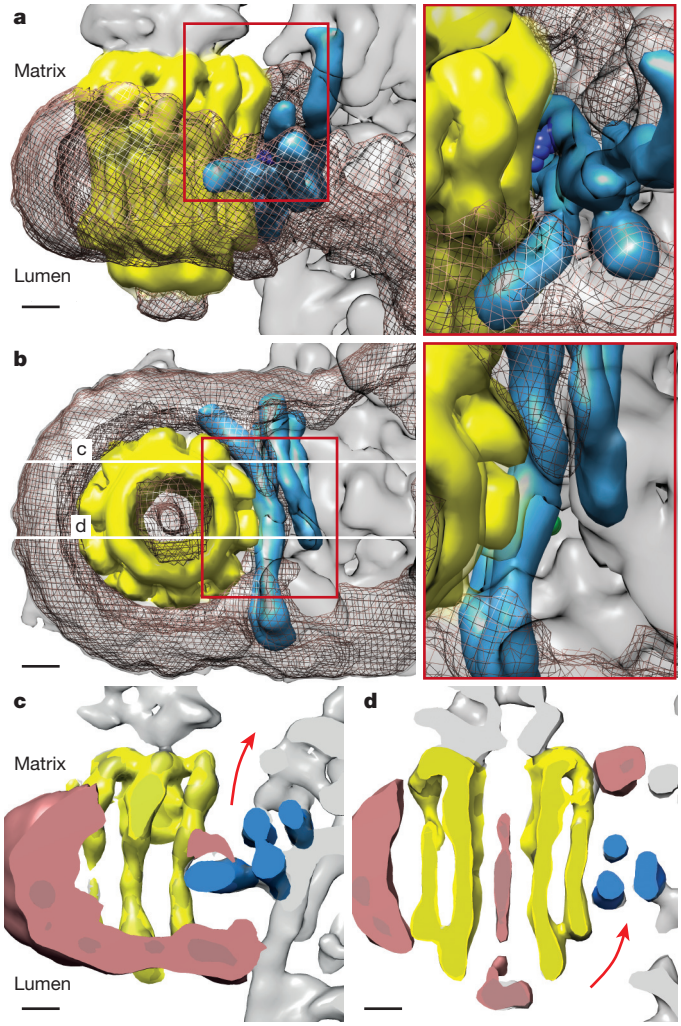


Figure 3 | Two aqueous half-channels. a, Left: side view of membrane domain in one protomer. Light blue, *a*-subunit helices; yellow, *c*-ring. The aqueous matrix cavity (boxed) is lined by amphipathic segments of the long hairpin helices and exposes the matrix ends of three *c*-subunits. Right: the aqueous matrix cavity narrows to a slit at the point where the *a*- and *c*-subunits interact (dark blue). **b**, Left: lumenal view of membrane region with aqueous cavity (boxed). Right: lumenal cavity with proposed position of interchangeable residue pair *a*E288/*a*H248 (green). Cross-sections along white lines in **b**, through the aqueous cavities on the matrix (**c**) and lumenal (**d**) side. Red arrows mark the proton paths into the lumenal (**d**) and out of the matrix (**c**) half-channels. Pink, detergent shell. Scale bar, 10 Å. See also Supplementary Video 3.

Structures of *c*-rings and molecular dynamics simulations suggest that the transition from a hydrophobic to a hydrophilic environment is sufficient to convert the proton-binding glutamic acid of the *c*-ring from a locked, protonated state to an open conformation, where the proton can be exchanged[7,19]. In ATP synthesis, the *c*-ring rotates anticlockwise as seen from the matrix. This means that a protonated *c*-ring glutamate encounters the C-terminal part of the longest *a*-subunit helix before the solvent channel. All compared C-terminal *a*-subunit sequences contain polar and charged residues in this region, which may locally disrupt the matrix membrane leaflet, forming the aqueous matrix cavity. The hydrophilic environment in this cavity would destabilize the locked conformation, and thus promote the release of the proton into the matrix.

Our model explains the direction in which the *c*-ring rotates under ATP synthesis conditions (Fig. 4). Thermal motion causes the ring to move backwards and forwards randomly as a Brownian ratchet[22]. A *c*-subunit glutamate accepts a proton through the lumenal half-channel, where the proton concentration is higher than on the matrix side. The

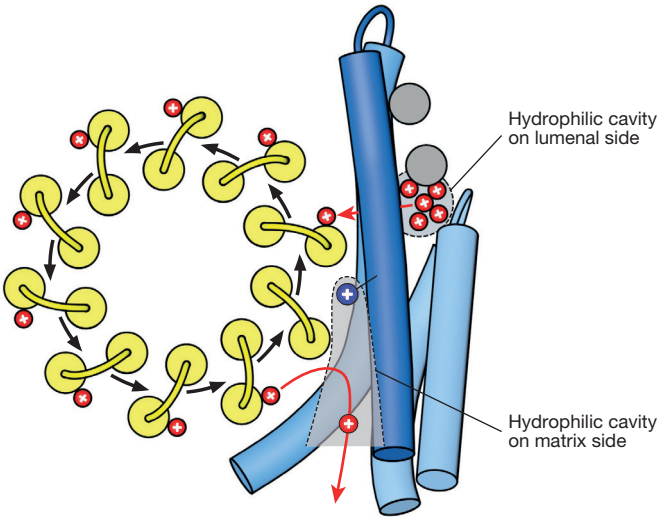


Figure 4 | Proton translocation through F-type ATP synthases. Protons (red) reach the conserved glutamate in the *c*-subunit via the aqueous lumenal half-channel (dashed grey circle). The proton competes with the strictly conserved *a*R239 (blue) for interaction with *c*-ring glutamates, which carry the proton around the *c*-ring. When the *c*-subunit approaches the hydrophilic half-channel on the matrix side (dashed grey outline), the glutamate becomes hydrated and adopts an open conformation, from which the proton can escape into the matrix. Solid grey circles indicate transmembrane helices.

protonated glutamate cannot pass the conserved arginine[4,23] but instead partitions into the hydrophobic environment of the lipid bilayer in a locked conformation[7,19]. On the other side of the membrane, the polar environment of the matrix cavity hydrates another protonated *c*-ring glutamate, which adopts an open conformation, so that the proton can escape to the high-pH matrix solvent. The de-protonated glutamate cannot partition back into the lipid phase but passes the conserved arginine to become re-protonated. Because of the relative locations of the two half-channels, the conserved arginine and the direction of the electrochemical gradient, the ring can rotate only one way, unless energy released by ATP hydrolysis drives it in the opposite direction. By this mechanism, the electrochemical membrane gradient is converted into the torque that drives ATP synthesis in mitochondria, bacteria and chloroplasts.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 August; accepted 29 December 2014.
Published online 23 February 2015.

1. Pogoryelov, D. *et al.* Engineering rotor ring stoichiometries in ATP synthases. *Proc. Natl Acad. Sci. USA* **109,** E1599–E1608 (2012).
2. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F_1 ATPase from bovine heart mitochondria. *Nature* **370,** 621–628 (1994).
3. Davies, K. M. *et al.* Macromolecular organization of ATP synthase and complex I in whole mitochondria. *Proc. Natl Acad. Sci. USA* **108,** 14121–14126 (2011).
4. Mitome, N. *et al.* Essential arginine residue of the F_o-*a* subunit in F_oF_1-ATP synthase has a role to prevent the proton shortcut without c-ring rotation in the F_o proton channel. *Biochem. J.* **430,** 171–177 (2010).
5. Cain, B. D. & Simoni, R. D. Interaction between Glu-219 and His-245 within the a-subunit of F_1F_o-ATPase in *Escherichia coli*. *J. Biol. Chem.* **263,** 6602–6612 (1988).
6. van Lis, R., Mendoza-Hernández, G., Groth, G. & Atteia, A. New insights into the unique structure of the F_0F_1-ATP synthase from the chlamydomonad algae *Polytomella* sp. and *Chlamydomonas reinhardtii*. *Plant Physiol.* **144,** 1190–1199 (2007).
7. Symersky, J. *et al.* Structure of the c(10) ring of the yeast mitochondrial ATP synthase in the open conformation. *Nature Struct. Mol. Biol.* **19,** 485–491 (2012).
8. Rees, D. M., Leslie, A. G. & Walker, J. E. The structure of the membrane extrinsic region of bovine ATP synthase. *Proc. Natl Acad. Sci. USA* **106,** 21597–21601 (2009).

9. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W. & Bork, P. Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309,** 1–18 (2001).
10. Jiang, W. & Fillingame, R. H. Interacting helical faces of subunits *a* and *c* in the F_1F_0 ATP synthase of *Escherichia coli* defined by disulfide cross-linking. *Proc. Natl Acad. Sci. USA* **95,** 6607–6612 (1998).
11. Moore, K. J. & Fillingame, R. H. Structural interactions between transmembrane helices 4 and 5 of subunit *a* and the subunit c ring of *Escherichia coli* ATP synthase. *J. Biol. Chem.* **283,** 31726–31735 (2008).
12. Hakulinen, J. K. *et al.* A structural study on the architecture of the bacterial ATP synthase F_o motor. *Proc. Natl Acad. Sci. USA* **109,** E2050–E2056 (2012).
13. Lau, W. C. Y. & Rubinstein, J. L. Subnanometre-resolution structure of the intact *Thermus thermophilus* H^+-driven ATP synthase. *Nature* **481,** 214–218 (2012).
14. Schwem, B. E. & Fillingame, R. H. Cross-linking between helices within subunit *a* of *Escherichia coli* ATP synthase defines the transmembrane packing of a four-helix bundle. *J. Biol. Chem.* **281,** 37861–37867 (2006).
15. Senes, A., Engel, D. E. & DeGrado, W. F. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **14,** 465–479 (2004).
16. Hoppe, J., Schairer, H. U., Friedl, P. & Sebald, W. An Asp-Asn substitution in the proteolipid subunit of the ATP-synthase from *Escherichia coli* leads to a non-functional proton channel. *FEBS Lett.* **145,** 21–29 (1982).
17. Hatch, L. P., Cox, G. B. & Howitt, S. M. The essential arginine residue at position 210 in the *a* subunit of the *Escherichia coli* ATP synthase can be transferred to position 252 with partial retention of activity. *J. Biol. Chem.* **270,** 29407–29412 (1995).
18. Angevine, C. A. & Fillingame, R. H. Aqueous access channels in subunit *a* of rotary ATP synthase. *J. Biol. Chem.* **278,** 6066–6074 (2003).
19. Pogoryelov, D. *et al.* Microscopic rotary mechanism of ion translocation in the F_o complex of ATP synthases. *Nature Chem. Biol.* **6,** 891–899 (2010).
20. Moore, K. J., Angevine, C. A., Vincent, O. D., Schwem, B. E. & Fillingame, R. H. The cytoplasmic loops of subunit *a* of *Escherichia coli* ATP synthase may participate in the proton translocating mechanism. *J. Biol. Chem.* **283,** 13044–13052 (2008).
21. Steed, P. R. & Fillingame, R. H. Residues in the polar loop of subunit c in *Escherichia coli* ATP synthase function in gating proton transport to the cytoplasm. *J. Biol. Chem.* **289,** 2127–2138 (2014).
22. Junge, W., Lill, H. & Engelbrecht, S. ATP synthase: an electrochemical transducer with rotatory mechanics. *Trends Biochem. Sci.* **22,** 420–423 (1997).
23. Matthies, D. *et al.* High-resolution structure and mechanism of an F/V-hybrid rotor ring in a Na^+-coupled ATP synthase. *Nature Commun.* **5,** http://dx.doi:10.1038/ncomms6286 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Meier and J. D. Faraldo-Gómez for discussions and reading the manuscript. Ö. Yildiz and J. F. Castillo-Hernandez provided computer support. This work was funded by the Max Planck Society (M.A., N.K., D.J.M., J.V., K.M.D., W.K.) and the Deutsche Forschungsgemeinschaft Cluster of Excellence Frankfurt 'Macromolecular Complexes' (K.M.D., W.K.).

Author Contributions K.M.D., M.A. and W.K. designed the experiments. N.K. purified the protein. M.A., K.M.D., N.K. and D.J.M. collected images. M.A., N.K., K.M.D., J.V. and D.J.M. processed data. K.M.D., M.A., J.V. and W.K. analysed the data and wrote the paper.

Author Information The electron cryomicroscopy map of the *Polytomella* sp. proton-driven ATP synthase dimer has been deposited in the Electron Microscopy Data Bank under accession number EMD-2852. Raw image data have been deposited in the European Bioinformatics Institute Electron Microscopy Pilot Image Archive (http://www.ebi.ac.uk/pdbe/emdb/empiar/) under accession number EMPIAR-10023. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.K. (werner.kuehlbrandt@biophys.mpg.de) or K.M.D. (karen.davies@biophys.mpg.de).

METHODS

Isolation of mitochondria. Cultures of *Polytomella* sp. (198.80, E. G. Pringsheim) were obtained from the Göttingen Collection of Algal Cultures (Sammlung Algenkulturen Göttingen (SAG), University of Göttingen, Germany) and grown at room temperature (23 ± 2 °C) without agitation in MAP medium[24] with the addition of trace elements and vitamins B_1 and B_{12} (ref. 25). Cultures were made axenic by serial dilution and the addition of carbenicillin (100 μg ml^{-1}). To isolate mitochondria, cells were harvested during their exponential growth phase (4 °C, 15 min, 10,500*g*), resuspended in a minimal volume of isolation buffer (350 mM sucrose, 20 mM MOPS/NaOH pH 7.4, 1 mM EGTA) supplemented with 0.3% (w/v) fat-free BSA and protease inhibitors (Mini cOmplete Protease Inhibitor Cocktail, Roche) and lysed by five passes through a ball-bearing homogenizer with 18 μm clearance, Isobiotec). The cell lysate was clarified by differential centrifugation, first at 2,000*g* to remove cell debris and then at 10,000*g* to pellet the mitochondria. Mitochondria were resuspended in minimal isolation buffer without BSA, frozen in liquid nitrogen and stored at −80 °C for later use.

Purification of ATP synthase dimers. *Polytomella* ATP synthase dimers were purified according to a published procedure[26] with modifications. Mitochondrial membranes (50 mg) were washed in buffer A (50 mM Tris-HCl, pH 8.0, 1 mM $MgCl_2$) and resuspended in solubilization buffer (50 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$, 3.33% (w/v) *n*-dodecyl-β-D-maltoside (DDM)) to a total volume of 1.5 ml. After 30 min incubation at 4 °C, membranes were removed by centrifugation and the supernatant was loaded onto a diethylaminoethyl (DEAE) anion exchange column (bed volume, 2.5 ml, DEAE-Biogel A, Biorad) equilibrated in buffer B (50 mM Tris-HCl, pH 8.0, 1 mM $MgCl_2$, 0.05% (w/v) DDM). The column was washed with buffer B-20, (buffer B with 20 mM NaCl) and the ATP synthase dimers were eluted with buffer B-55 (buffer B with 55 mM NaCl). The eluate was concentrated to 50 μl (Vivaspin 500 maximum spin speed columns with 50,000 molecular mass cutoff) and loaded onto a Superose 6 gel-filtration column (PC 3.2/30) equilibrated in buffer B-20 on an Ettan purifier (GE Healthcare). Fractions eluting at retention volume 1.05–1.15 ml were collected and analysed for ATPase activity[27].

Electron microscopy. A 3 μl aliquot of a 0.5 mg ml^{-1} ATP synthase dimer solution in 0.05% (w/v) DDM showing oligomycin and DCCD-sensitive ATPase activity was applied to R2/2 Quantifoil grids without carbon backing, and plunge-frozen in a Vitrobot (FEI) at 75% humidity, 10 °C with 9 s blotting time. Dose-fractionated 1.5 s movies of 24 frames with an electron dose of three electrons per square ångström per frame were recorded after coma-free alignment[28] with an FEI Polara electron microscope at 300 kV and 1.5–5 μm defocus on a back-thinned Falcon-II direct electron detector (FEI) as described[29]. Images were collected at a nominal magnification of ×59,000, corresponding to a specimen pixel size of 1.77 Å.

Image processing and map interpretation. Images without obvious drift were selected for further processing. Global beam induced motion of image frames was corrected[30]. The contrast transfer function for each image was determined using CTFFIND3 (ref. 31) within the RELION-1.3 workflow[32,33] and 49,600 particles were picked manually using EMAN boxer[34].

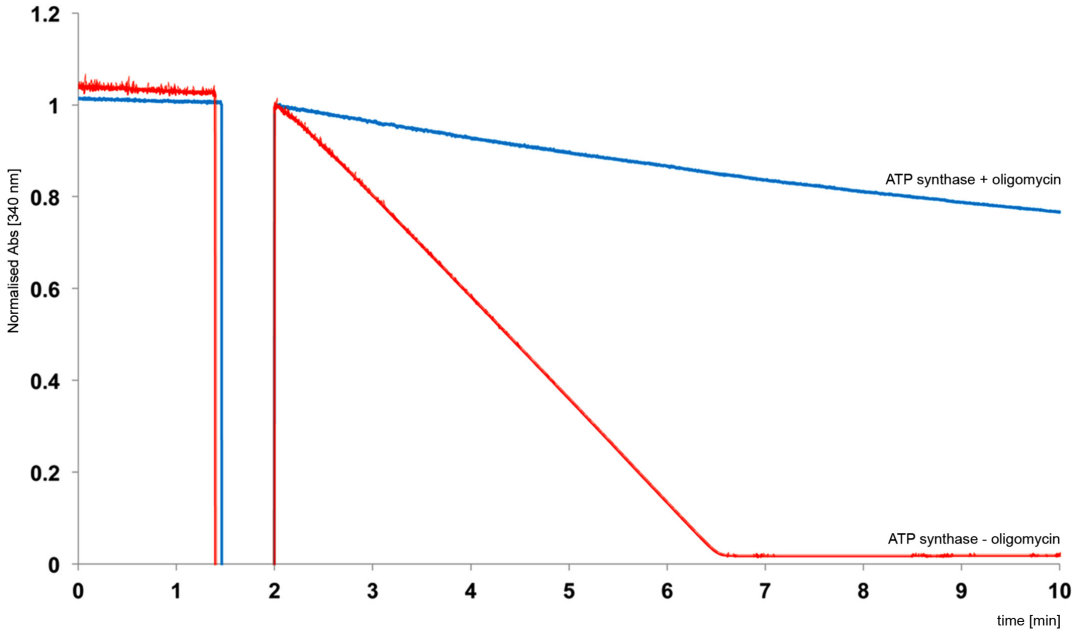
A sub-tomogram average of the *Polytomella* ATP synthase dimer, obtained from tomographic volumes of mitochondrial membranes, was used as an initial reference for refinement in RELION-1.3. The handedness of the sub-tomogram average was determined and confirmed by comparison with known X-ray structures of ATP synthase subcomplexes[35].

Two-dimensional classification of the picked particles was performed in RELION-1.3. Three-dimensional classification was performed to identify structural heterogeneity. Three out of the five classes obtained were virtually identical and showed well-resolved structural features. A total of 37,238 particles belonging to these three classes were combined for further processing. Individual frames were B-factor weighted and movements of individual particles were reversed by movie frame correction in RELION-1.3 (ref. 36). The final *c2*-averaged volume was calculated from particles in frames 1–15. A B-factor of −200 Å^2 for map sharpening was determined using the modulation transfer function of the Falcon II detector[37].

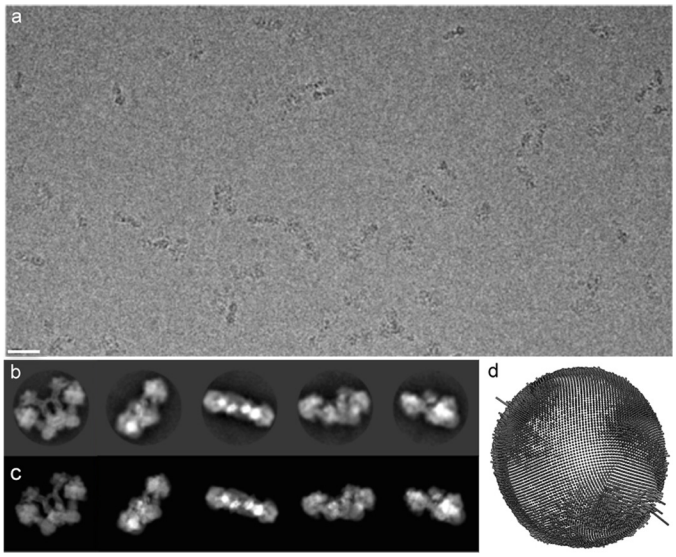
Resolution anisotropy was assessed using the local resolution estimation programme ResMap[38] (Extended Data Fig. 3a). Gold-standard Fourier shell correlations were calculated from two independently refined data sets to determine the resolution of the complete volume and sub-areas of the structure. For the latter, different regions of two independently refined maps were soft-masked and the resolution determined by Fourier shell correlations using EMAN2 (ref. 39) (Extended Data Fig. 3b).

Sequences were aligned with ClustalW2 (ref. 40). Helices were fitted with DireX[41] and atomic models with Chimera[42].

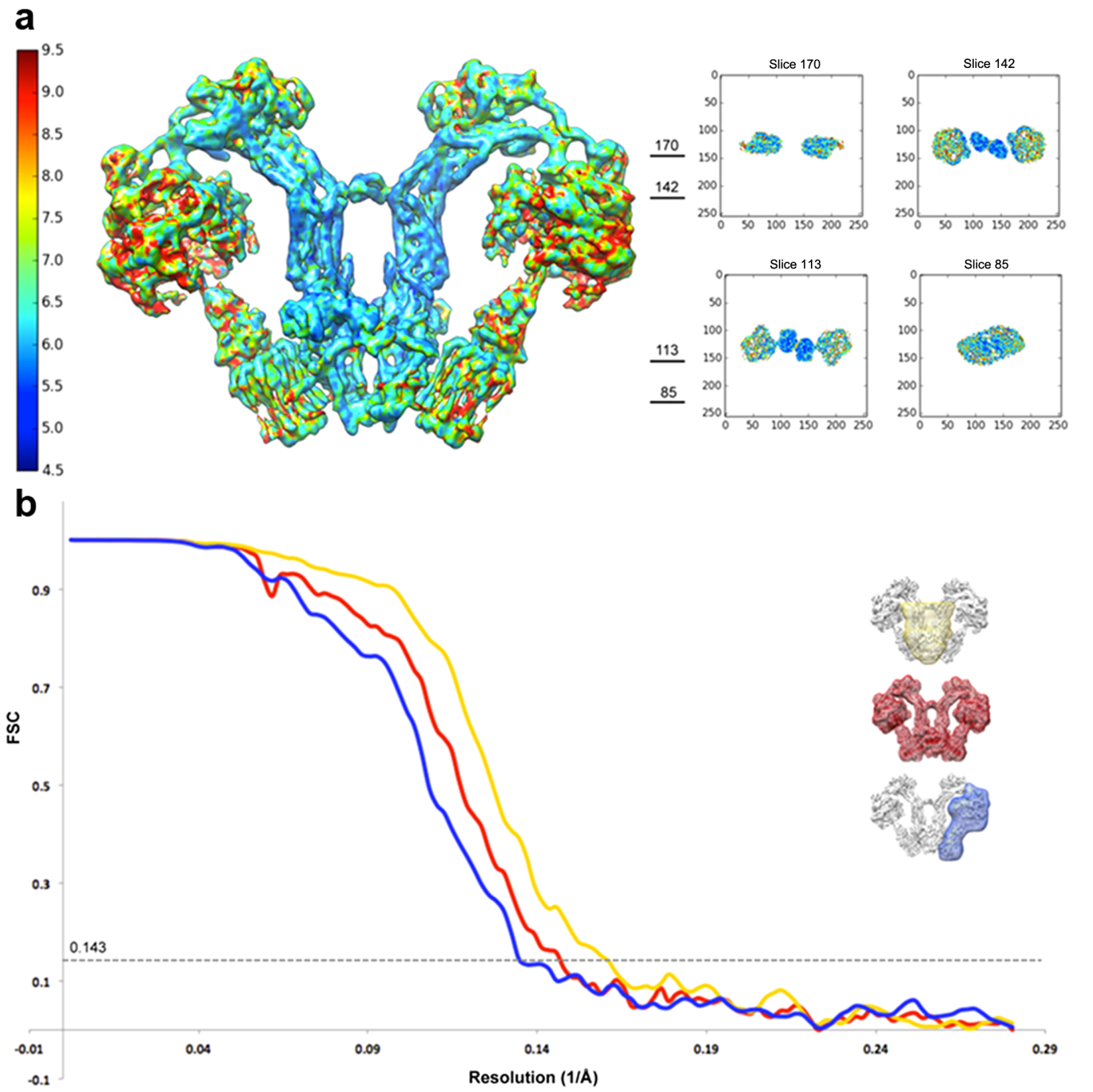
24. van Lis, R., González-Halphen, D. & Atteia, A. Divergence of the mitochondrial electron transport chains from the green alga *Chlamydomonas reinhardtii* and its colorless close relative *Polytomella*. *Biochim. Biophys. Acta* **1708,** 23–34 (2005).
25. Atteia, A., van Lis, R., Ramírez, J. & González-Halphen, D. *Polytomella* spp. growth on ethanol: extracellular pH affects the accumulation of mitochondrial cytochrome c_{550}. *Eur. J. Biochem.* **267,** 2850–2858 (2000).
26. Vázquez-Acevedo, M. *et al.* The mitochondrial ATP synthase of chlorophycean algae contains eight subunits of unknown origin involved in the formation of an atypical stator-stalk and in the dimerization of the complex. *J. Bioenerg. Biomembr.* **38,** 271–282 (2006).
27. Villavicencio-Queijeiro, A. *et al.* The fully-active and structurally-stable form of the mitochondrial ATP synthase of *Polytomella* sp. is dimeric. *J. Bioenerg. Biomembr.* **41,** 1–13 (2009).
28. Mills, D. J., Vitt, S., Strauss, M., Shima, S. & Vonck, J. De novo modeling of the F_{420}-reducing [NiFe]-hydrogenase from a methanogenic archaeon by cryo-electron microscopy. *eLife* **2,** e00218 (2013).
29. Allegretti, M., Mills, D. J., McMullan, G., Kühlbrandt, W. & Vonck, J. Atomic model of the F_{420}-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. *eLife* **3,** e01963 (2014).
30. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
31. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
32. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
33. Wong, W. *et al.* Cryo-EM structure of the *Plasmodium falciparum* 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3,** e3080 (2014).
34. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
35. Davies, K. M., Anselmi, C., Wittig, I., Faraldo-Gomez, J. D. & Kuhlbrandt, W. Structure of the yeast F_1F_o-ATP synthase dimer and its role in shaping the mitochondrial cristae. *Proc. Natl Acad. Sci. USA* **109,** 13602–13607 (2012).
36. Scheres, S. H. W. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* **3,** e03665 (2014).
37. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).
38. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
39. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
40. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
41. Schroeder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15,** 1630–1641 (2007).
42. Pettersen, E. F. *et al.* UCSF Chimera: a visualisation system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
43. Wada, T., Long, J. C., Zhang, D. & Vik, S. B. A novel labeling approach supports the five-transmembrane model of subunit *a* of the *Escherichia coli* ATP synthase. *J. Biol. Chem.* **274,** 17353–17357 (1999).
44. Careaga, C. L. & Falke, J. J. Thermal motions of surface alpha-helices in the D-galactose chemosensory receptor: detection by disulfide trapping. *J. Mol. Biol.* **226,** 1219–1235 (1992).
45. Vincent, O. D., Schwem, B. E., Steed, P. R., Jiang, W. & Fillingame, R. H. Fluidity of structure and swiveling of helices in the subunit c ring of *Escherichia coli* ATP synthase as revealed by cysteine-cysteine cross-linking. *J. Biol. Chem.* **282,** 33788–33794 (2007).



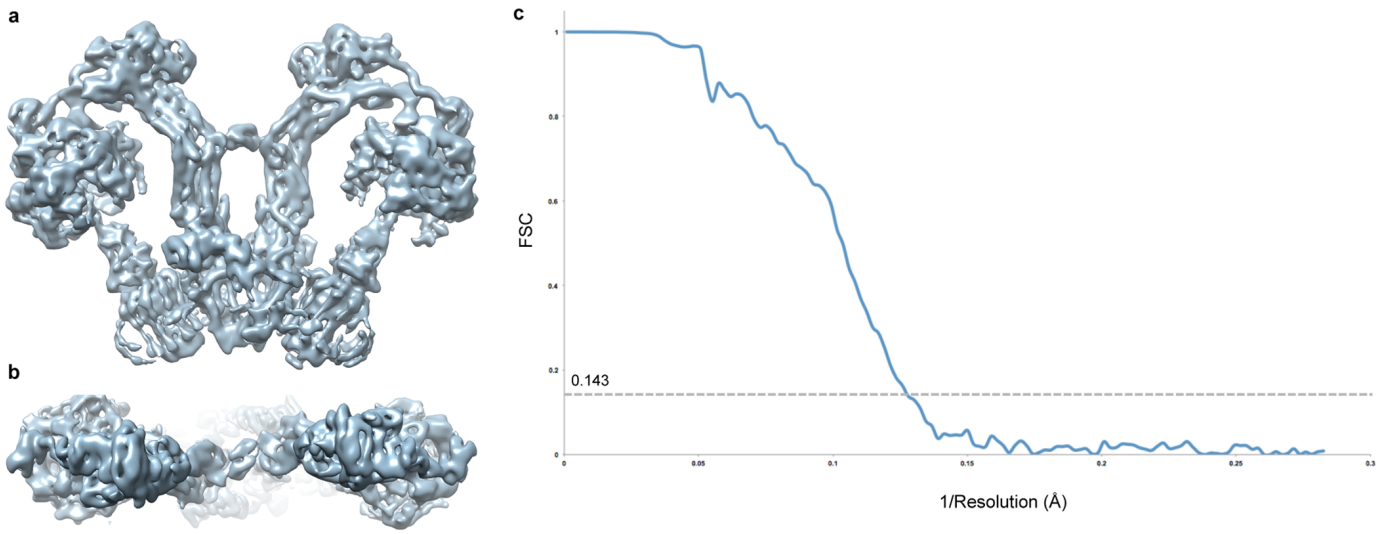
Extended Data Figure 1 | Oligomycin-sensitive ATPase activity. ATPase assays[27] performed with the *Polytomella* ATP synthase dimers used for electron cryomicroscopy data collection indicated high activity of 6–8 units per milligram protein, close to the reported activity for this complex[27]. ATPase activity was ~90% inhibited by oligomycin, indicating that the F_1F_o complex is coupled. The vertical lines indicate addition of dodecyl maltoside detergent to initiate the reaction. Measurements were performed in triplicates for each purification.



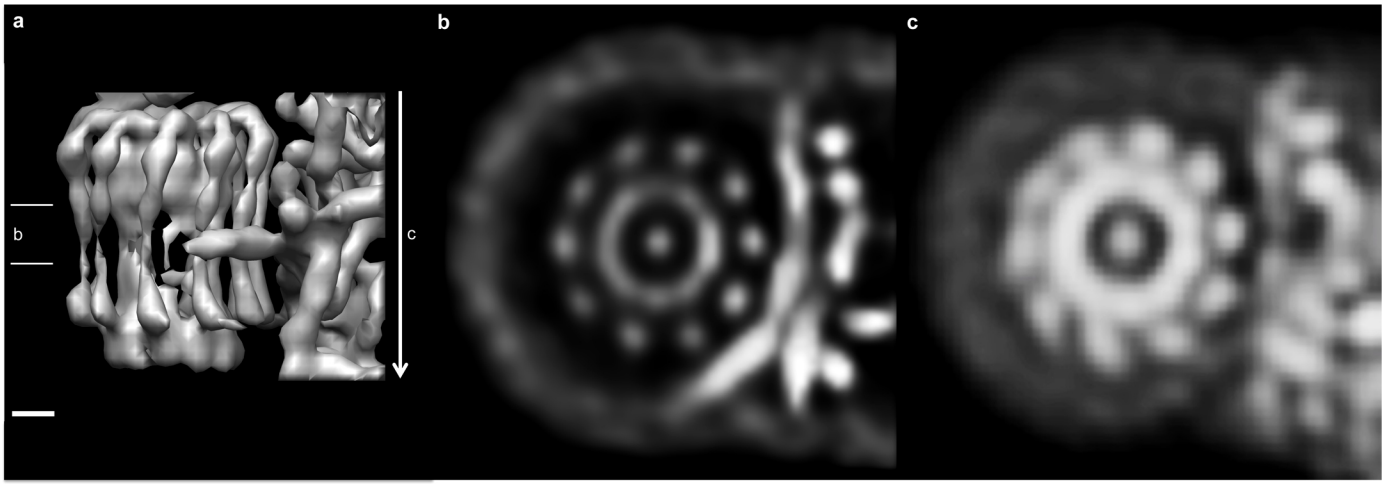
Extended Data Figure 2 | Projection images of the *Polytomella* F-type ATP synthase dimer. **a**, Typical electron cryo-micrograph of dimers in vitrified buffer recorded at 2.5 µm underfocus. **b**, Two-dimensional class averages and (**c**) corresponding projection images calculated from the final three-dimensional volume. **d**, Angular distribution of projection images used for three-dimensional reconstruction. ATP synthase dimers are oriented randomly in the thin layer of vitrified buffer. Scale bar in **a**, 50 nm.



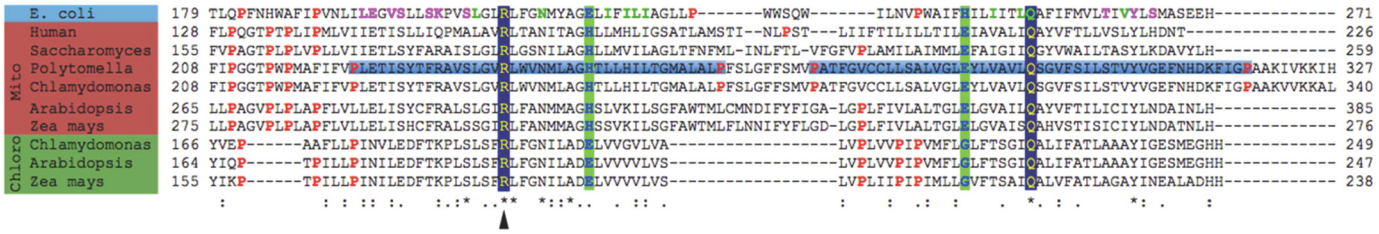
Extended Data Figure 3 | Local resolution estimates. **a**, The unfiltered electron cryomicroscopy map was analysed with the programme ResMap[38] to assess local resolution. The catalytic F_1 head (red to green) is less well-ordered than the peripheral stalk and the membrane-embedded *a*-subunit (green to blue), as indicated by the rainbow colour code on the left. The insets show cross-sections through the density at the levels indicated on the right. **b**, Fourier shell correlation curves calculated from two independently refined data sets after soft masking. The resolution was 7.0 Å for the whole complex (red), 7.4 Å for the F_1/*c*-ring complex (blue mask and curve) and 6.2 Å for the peripheral stalk and *a*-subunit (yellow mask and curve).



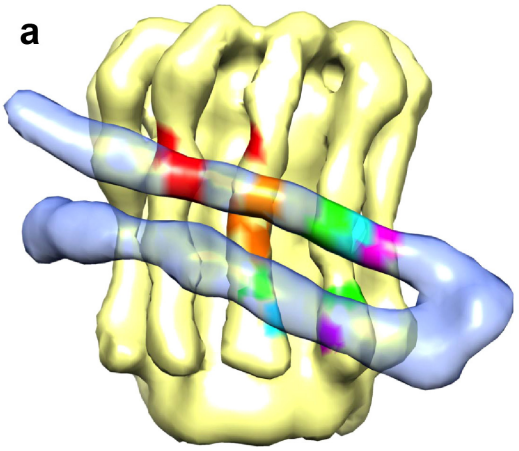
Extended Data Figure 4 | Resolution of unsymmetrized map. Side view (**a**) and top view (**b**) of the dimer map refined without imposing *c*2 symmetry (blue). **c**, The gold-standard Fourier shell correlation curve indicates a resolution of 8.0 Å for the unsymmetrized map, which shows all essential features of the *Polytomella* dimer, including *c2* symmetry.



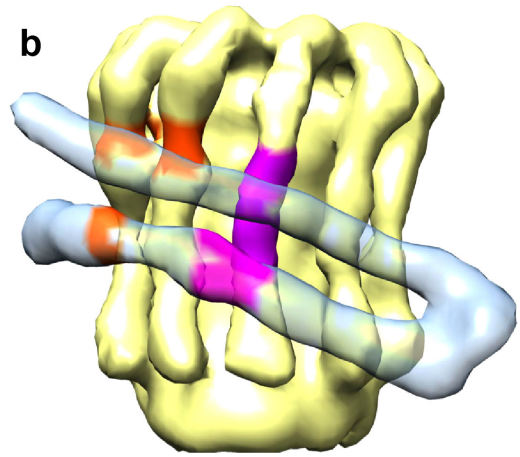
Extended Data Figure 5 | Horizontal helices in the three-dimensional map and in two-dimensional projection. **a**, Section of the three-dimensional map showing a side view of the c_{10}-ring on the left and the *a*-subunit on the right. **b**, A slice of the three-dimensional map volume indicated in **a** shows the horizontal helices clearly. **c**, In a two-dimensional map generated by projecting the three-dimensional map volume in **a** along the indicated direction (arrow), the long horizontal helices are in effect invisible, whereas the vertical *c*-ring helices stand out clearly. Scale bar, 10 Å.



Extended Data Figure 6 | Sequence alignment of about 120 C-terminal subunit *a* residues. Sequences of bacterial (blue), mitochondrial (red) and chloroplast (green) F-type ATP synthases are compared. The interchangeable residue pairs *a*R239/*a*Q295 and *a*E288/*a*H248 are dark blue and green, respectively; sequences of helices fitted in Fig. 2b are light blue. The black arrowhead marks the strictly conserved arginine essential for coupled proton translocation. Red, prolines; magenta, solvent-exposed residues[18,43]; green, residues that crosslink to the *c*-ring[10,11]. Identical (*) or similar ('.' or ':') residues are indicated.



	E. coli residues	Cβ distance in model (Å)	*Polytomella* residues
	a207/c55	10	a236/c53
	a214/c62	9	a243/c60
	a214/c65	10	a243/c63
	a221/c69	14	a250/c67
	a223/c72	16	a252/c70
	a224/c73	14	a253/c71
	a225/c73	19	a254/c71



	E. coli residues	Cβ distance in model (Å)	*Polytomella* residues
	a263/c54	15	a306/c52
	a262/c53	23	a305/c51
	a262/c54	18	a305/c52
	a262/c55	17	a305/c53
	a252/c62	8	a295/c60
	a252/c65	9	a295/c63
	a251/c57	21	a294/c55
	a251/c59	17	a294/c57
	a251/c62	13	a294/c60
	a248/c62	16	a291/c60
	a248/c63	12	a291/c61
	a248/c65	10	a291/c63

Extended Data Figure 7 | Subunit *a* and *c* crosslinking distances. Pairs of *a*- and *c*-subunit residues crosslinked in *E. coli* using (**a**) zero-length crosslinkers[10] and (**b**) bis-MTS reagents[11]. Tabulated distances were measured between beta carbons of residues in the modelled *a*-subunit helices and *Saccharomyces c*-ring helices fitted to the map. Zero-length crosslinks can be considerably shorter than crystallographic distances[44,45], most probably as a result of thermal motion during the long incubation times required for crosslinking[10]. Nevertheless, the inter-residue distances in our model agree with the *E. coli* crosslinking data within reasonable margins, allowing for species differences between bacterial and mitochondrial ATP synthases.

LETTER

doi:10.1038/nature14365

Electron cryomicroscopy observation of rotational states in a eukaryotic V–ATPase

Jianhua Zhao[1,2]*, Samir Benlekbir[1]* & John L. Rubinstein[1,2,3]

Eukaryotic vacuolar H^+-ATPases (V-ATPases) are rotary enzymes that use energy from hydrolysis of ATP to ADP to pump protons across membranes and control the pH of many intracellular compartments. ATP hydrolysis in the soluble catalytic region of the enzyme is coupled to proton translocation through the membrane-bound region by rotation of a central rotor subcomplex, with peripheral stalks preventing the entire membrane-bound region from turning with the rotor. The eukaryotic V-ATPase is the most complex rotary ATPase: it has three peripheral stalks, a hetero-oligomeric proton-conducting proteolipid ring, several subunits not found in other rotary ATPases, and is regulated by reversible dissociation of its catalytic and proton-conducting regions[1,2]. Studies of ATP synthases, V-ATPases, and bacterial/archaeal V/A-ATPases have suggested that flexibility is necessary for the catalytic mechanism of rotary ATPases[3–5], but the structures of different rotational states have never been observed experimentally. Here we use electron cryomicroscopy to obtain structures for three rotational states of the V-ATPase from the yeast *Saccharomyces cerevisiae*. The resulting series of structures shows ten proteolipid subunits in the c-ring, setting the ATP:H^+ ratio for proton pumping by the V-ATPase at 3:10, and reveals long and highly tilted transmembrane α-helices in the a-subunit that interact with the c-ring. The three different maps reveal the conformational changes that occur to couple rotation in the symmetry-mismatched soluble catalytic region to the membrane-bound proton-translocating region. Almost all of the subunits of the enzyme undergo conformational changes during the transitions between these three rotational states. The structures of these states provide direct evidence that deformation during rotation enables the smooth transmission of power through rotary ATPases.

The V-ATPase from *S. cerevisiae* consists of subunits $A_3B_3CD E_3FG_3Hac_xc'_yc''_zde$ (Extended Data Fig. 1), where *x*, *y*, and *z* denote unknown stoichiometries. In the soluble V_1 region, the three pairs of A- and B-subunits are arranged around the D-subunit of the central rotor. Each AB pair contains a single catalytic nucleotide-binding site and together the three AB pairs produce a pseudo-symmetric trimer of heterodimers, with each AB pair in a different conformation[6]. The three main conformations of the AB pairs are known as 'tight', 'loose', and 'open'. These conformations are expected to bind ATP, ADP and phosphate, and no nucleotide, respectively, although their nucleotide content in crystal structures of catalytic regions from other rotary ATPases has depended on the crystallization conditions[7,8]. During catalysis, the inter-conversion between the catalytic conformations of the three AB pairs is coupled to rotation of the central rotor, which extends into the membrane via subunit d, to the $c_xc'_yc''_z$-ring. This rotary catalytic mechanism suggests that a population of intact V-ATPase complexes will be found in three or more structurally distinct rotational states. To investigate the rotational states of the eukaryotic V-ATPase, we isolated the complex from *S. cerevisiae* in detergent and imaged it by electron cryomicroscopy (cryo-EM) with a direct detector device camera (Extended Data Fig. 2). Classification of V-ATPase images enabled the identification of three distinct three-dimensional classes of varying size that gave maps at resolutions of 6.9 Å (47% of images), 7.6 Å (36% of images), and 8.3 Å (17% of images) (Extended Data Fig. 3a, b), with 106,445 particle images ultimately contributing to

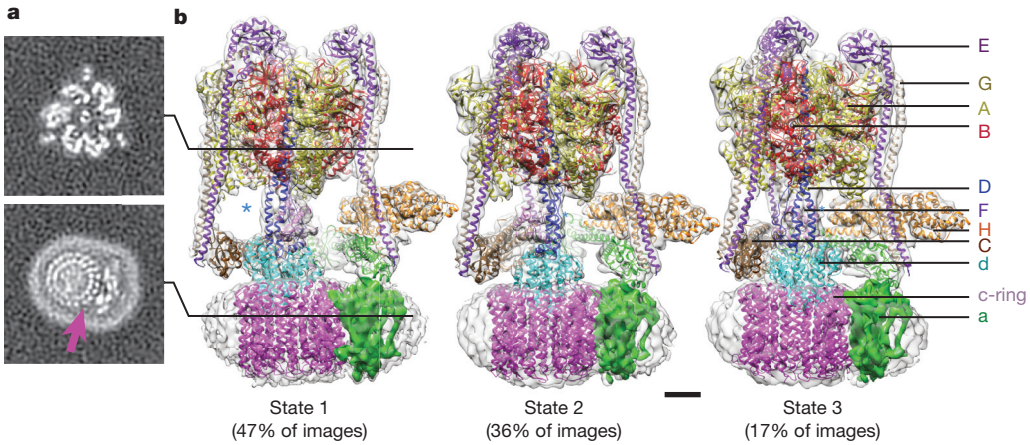


Figure 1 | Rotational states of the V-ATPase. a, Cross-sections through a three-dimensional map of rotational state 1 of the V-ATPase V_1 region (upper) and V_O region (lower) show that α-helices can be resolved. The best resolution in the V_O region is indicated with the pink arrow. **b**, The three different maps each allow docking of atomic models of subunits. The experimental map is shown in semi-transparent grey. The blue stars indicate the central rotor. Scale bar, 25 Å.

[1]Molecular Structure and Function Program, The Hospital for Sick Children Research Institute, 686 Bay Street, Toronto, Ontario M5G 0A4, Canada. [2]Department of Medical Biophysics, The University of Toronto, Toronto Medical Discovery Tower, MaRS Centre, 101 College Street, Toronto, Ontario M5G 1L7, Canada. [3]Department of Biochemistry, The University of Toronto, 1 King's College Circle, Medical Sciences Building, Toronto, Ontario M5S 1A8, Canada.
*These authors contributed equally to this work.

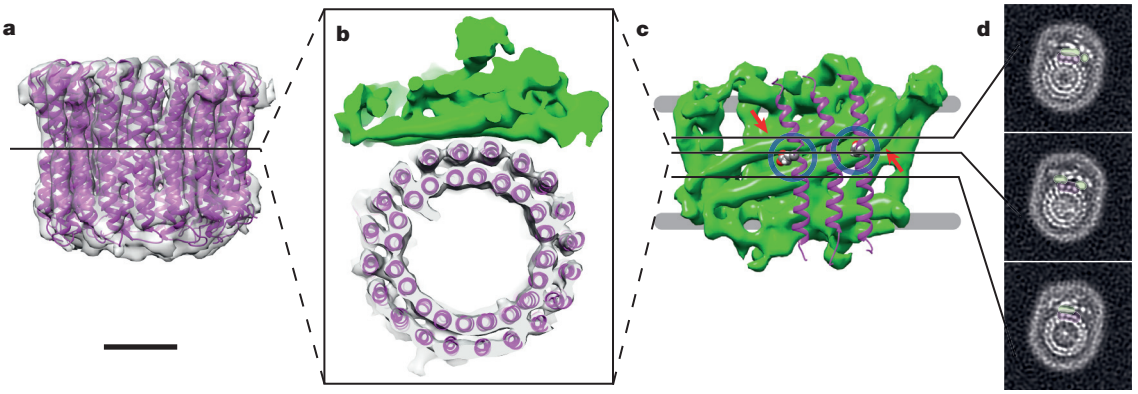


Figure 2 | The membrane-bound V_O region. **a**, The c-ring segment from the cryo-EM map accommodates ten c-, c′-, or c′′-subunits (magenta). **b**, The c-ring sits adjacent to the a-subunit (green). **c**, The a-subunit contains two long and highly tilted α-helices (red arrows). The conserved Glu residues of the c-subunits lie between the two tilted α-helices of the a-subunit. In the eukaryotic V-ATPase c-ring, every other outer α-helix has a Glu residue (blue circles). Grey lines indicate the approximate membrane boundaries. **d**, The tilted α-helices of the a-subunit (green) contact three different α-helices of the c-ring (magenta). Scale bar, 25 Å.

the three maps. Cross-sections through the maps show that α-helices are well resolved at these resolutions (Fig. 1 and Supplementary Video 1). The class with the largest population most closely corresponds to an earlier cryo-EM map of the *S. cerevisiae* V-ATPase where conformational separation was not performed[9] but differs from the conformation identified for the *Manduca sexta* V-ATPase[10].

Resolution does not appear to be homogenous within each three-dimensional map, suggesting further conformational heterogeneity within each class. However, additional classification strategies involving focusing classification on the regions thought to be variable[11] did not produce meaningful classes. The resolution appears better in the soluble V_1 region of the complex (Fig. 1a top and Extended Data Fig. 3c) than in the membrane-bound V_O region (Fig. 1a bottom and Extended Data Fig. 3c). In the V_O region resolution is best where the a-subunit interacts with the c-ring (pink arrow in Fig. 1a bottom), suggesting a rigid interaction within an otherwise dynamic complex. Crystal structures and homology models for V-ATPase subunits were docked into the three different maps and their conformations refined by molecular dynamics flexible fitting[12] (Fig. 1b and Extended Data Fig. 4). The most striking structural difference between the three maps is the position of the central rotor, consisting of subunits D, F, d, and the c-ring (blue star in Fig. 1b). Each map therefore appears to correspond to a rotational state of the enzyme. Rotational state 3 is the most different from the other two and is the class with the fewest particle images. The unequal distribution of particle images in the three classes suggests that, in the absence of free nucleotide, the V-ATPase relaxes to three unequally populated states.

The proton-carrying c-ring of the yeast V-ATPase is a hetero-oligomer of subunits c, c′, and c′′, each possessing a single conserved glutamate residue (Glu137, Glu145, Glu108, respectively) that can bind and transport protons during catalysis[13]. Subunits c and c′ have four transmembrane α-helices each, with an additional amino (N)-terminal α-helix for c′′ that may not be membrane bound and is not necessary for function[14]. The cryo-EM maps show the c-ring to consist of an inner ring and an outer ring of transmembrane α-helices (Fig. 1a bottom and Fig. 2a, b magenta). The ring appears to have 20-fold symmetry (Fig. 1a bottom and Fig. 2b), suggesting that the extra N-terminal α-helix of the c′′-subunit may protrude into the ring, as seen in a recent crystal structure of a bacterial heteromeric c-ring[15]. The structure of the c-ring in each cryo-EM map accommodates a total of ten c-, c′-, or c′′-subunits (Fig. 2a, b) with each subunit contributing two α-helices to the outer ring and two α-helices to the inner ring. The modelled c-subunits fit the density with their N and carboxy (C) termini facing the luminal side of the membrane[16]. The presence of ten subunits in the c-ring is inconsistent with an earlier prediction of a 4:1:1 stoichiometry for subunits c, c′, and c′′, which would require a total of 6 or 12 subunits in the ring[17,18]. A result of the existence of ten subunits in the c-ring is that in a tightly coupled enzyme complete rotation of the c-ring driven by hydrolysis of three ATP molecules would deliver ten protons across the lipid bilayer. Consequently, the c-ring stoichiometry probably sets the ATP:H^+ ratio at 3:10 for the *S. cerevisiae* V-ATPase, which is the same ratio found with the F-type ATP synthase in *S. cerevisiae* mitochondria[19]. The ATP:H^+ ratio was not known previously for any eukaryotic V-ATPase. With a Gibbs free energy for ATP hydrolysis of 57 kJ mol^{-1} at 30 °C, this ATP:H^+ ratio limits the maximum pH gradient or voltage established across the vacuolar membrane in *S. cerevisiae* to 3.0 units or 180 mV, respectively[20].

The three maps provide the highest-resolution insight available into the structure of the membrane-bound portion of subunit a, which contains the channels that conduct protons to and from the c-ring. The transmembrane proteins in the V_O region of the *S. cerevisiae* V-ATPase are subunits a, e, and the c-ring. However, the dodecylmaltoside-solubilized V-ATPase does not contain the e-subunit[21]. Consequently, any density in the V_O region that cannot be attributed to the c-ring or detergent must be part of the a-subunit (Fig. 2b–d). Subunit a has been predicted to have eight transmembrane α-helices[22]. The density corresponding to the a-subunit has a complex fold and appears to have at least eight transmembrane α-helices and several well-defined structural elements above and below the expected position of the lipid bilayer. Strikingly, there are two highly tilted α-helices from the a-subunit that span the lipid bilayer where the a-subunit is in contact with the c-ring (Fig. 2c, red arrows). These α-helices contact a group of three α-helices from the c-ring (Fig. 2c, d). In F-type ATP synthases and V/A-ATPases each outer α-helix of the c-ring has a conserved proton-carrying Glu or Asp residue. In the eukaryotic V-ATPase every other α-helix on the surface of the c-ring has the conserved Glu residue. At the present resolution it is not possible to distinguish the outer helices with the conserved Glu residues from the outer helices lacking the Glu residues. Therefore, it is possible to fit the c-ring in the electron microscopy map in two different ways: one where a single Glu residue contacts the a-subunit and one where two different Glu residues contact the a-subunit. This latter arrangement would place two different c-subunits (Fig. 2c, circled in blue) in different chemical environments, enabling one c-subunit to exchange protons with the cytoplasm while the other exchanges protons with the organelle matrix. This arrangement could create two half-channels for proton translocation[23] and was seen with the *Thermus thermophilus* enzyme[24]. The α-helices in the a-subunit are much easier to identify in the present maps at 6.9–8.3 Å resolution than in the map of the *T. thermophilus* V/A-ATPase at 9.7 Å resolution determined earlier[24]. The α-helices identified here in all three states (Extended Data Fig. 5a) are consistent with the density from the earlier map of the bacterial enzyme (Extended Data Fig. 5b left) but are

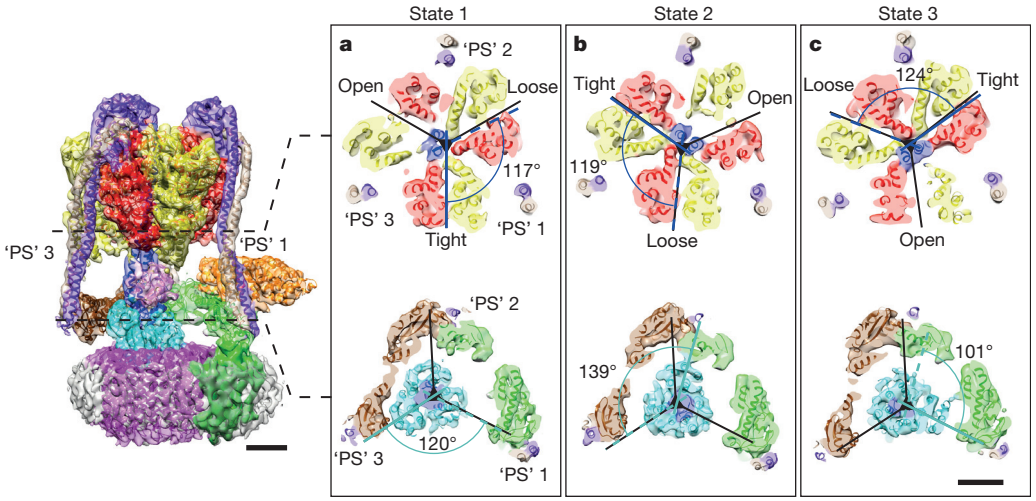


Figure 3 | Symmetry mismatch between the V_1 and V_O regions. Upper, the central rotor D-subunit (dark blue) makes ~120° steps (dashed blue line to blue line) from state 3 to 1 (**a**), 1 to 2 (**b**), and 2 to 3 (**c**) relative to the rest of the V_1 region, which itself undergoes a slight twist relative to the V_O region (black lines). Lower, in the V_O region, the d-subunit and c-ring undergo rotations of ~120°, ~139°, and ~101° from state 3 to 1 (**a**), 1 to 2 (**b**), and 2 to 3 (**c**), respectively (dashed cyan line to cyan line). PS, peripheral stalk. Scale bar, 25 Å.

not entirely consistent with the previously proposed locations of α-helices in that density (Extended Data Fig. 5b right).

With three catalytic nucleotide-binding sites, the V_1 region is expected to operate as a three-step motor while the ten titratable Glu residues of the c-ring suggest that the V_O region functions as a ten-step motor. This 3:10 ratio produces a symmetry mismatch between the V_1 and V_O regions. The new maps demonstrate how symmetry mismatch can be tolerated in rotary ATPases. The rotational position of subunit D of the central rotor was measured relative to the A_3B_3 hexamer (Fig. 3 upper, blue density and line), which itself twists within the rest of the enzyme between the three different states (Fig. 3, black lines). The rotation of subunit D corresponds to steps of 117° from state 3 to 1, 119° from state 1 to 2, and 124° from state 2 to 3, all of which are in good agreement with the 120° steps expected for the threefold

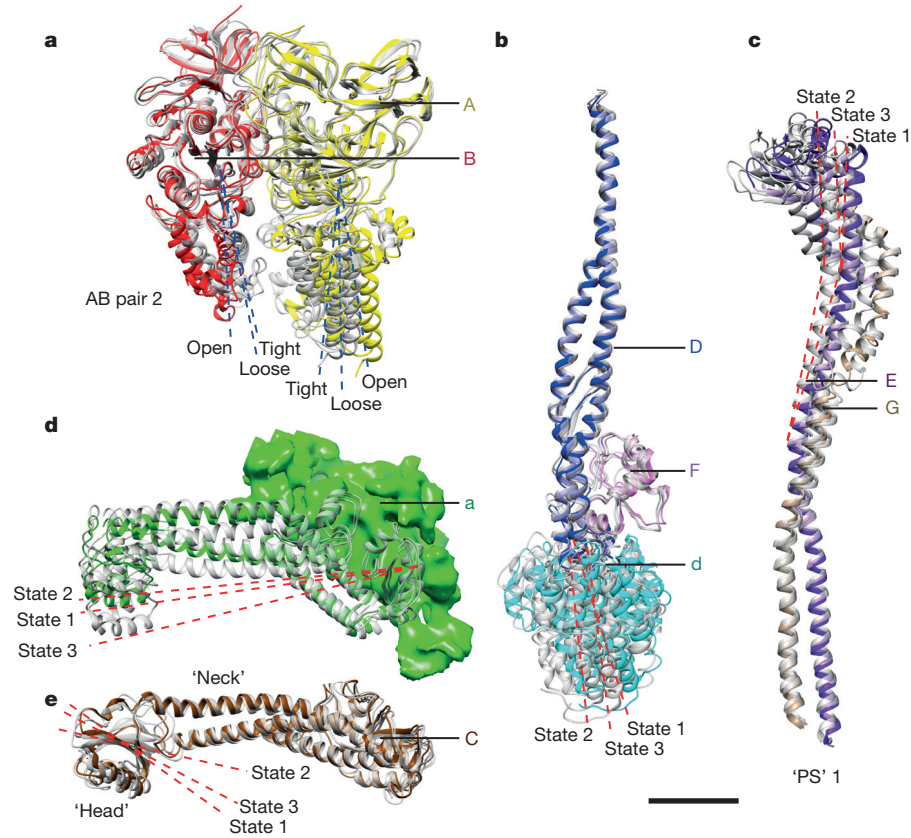


Figure 4 | Flexibility of subunits. **a**, Comparison of AB pair 2 from the rotational states (state 2 in colour, states 1 and 3 in grey) shows that the pair progresses through 'open', 'tight', and 'loose' conformations. **b**–**e**, Overlay of structures from the rotational states illustrates that during rotation the d-subunit wobbles relative to the DF subcomplex (**b**), the peripheral stalks bend (**c**), the N-terminal domain of subunit a pivots towards the central rotor (**d**), and the 'head' domain of the C-subunit twists relative to the 'neck' (**e**). These movements are most apparent in Supplementary Video 3. Scale bar, 25 Å.

pseudo-symmetric V_1 region. In the V_O region, rotation of one c-subunit against the a-subunit requires a 36° rotation of the c-ring (360°/10). The rotational position of the d-subunit and c-ring relative to the a-subunit could be measured precisely in rotational states 1 and 2 owing to the resolution in the membrane region of these maps. In rotational state 3, the rotation of the c-ring against the a-subunit could not be measured with the same confidence. For the transition from state 1 to 2, we measured a 139° rotation of the d-subunit and c-ring relative to the transmembrane portion of the a-subunit (Fig. 3 lower, cyan density and line). This rotation from state 1 to 2 matches the 144° rotation (36° × 4) expected from a four-tenths rotation of the c-ring and would be expected to deliver four protons across the lipid bilayer. The transitions from states 2 to 3 and 3 to 1 were measured at ~101° and ~120°, respectively. Together, these rotations match ~216° of rotation (36° × 6), and are probably due to 108° (36° × 3) of rotation for each transition. Consequently, it appears that transitions from states 1 to 2, 2 to 3, and 3 to 1 transport four, three, and three protons, respectively. The observed rotational states reveal the conformations of the enzyme *in vitro* after the available ATP in solution has been hydrolysed. Therefore, it is possible that the conformations of the enzyme when rapidly hydrolysing ATP could be different from the conformations observed here. For example, when rapidly hydrolysing ATP, the stepping motions of the enzyme may be smoothed with 3.3 protons transported for each ATP hydrolysis event. It is also possible that some 'slip' occurs during proton translocation and on average less than one proton is transported for each c-subunit.

Overlaying the different conformations of various subunits in the complex suggests some of the structural changes that may occur during rotary catalysis (Fig. 4). These conformational changes are illustrated dramatically by interpolating between the three rotational state structures (Supplementary Video 2). The different conformations of the catalytic subunits of rotary ATPases have been reported from mitochondrial and bacterial F_1-ATPases[25,26] and bacterial V_1/A_1-ATPases[8,27] but never before for a eukaryotic V-type enzyme and never before within an intact rotary ATPase. Different from crystal structures of isolated V_1/A_1-ATPase or F_1-ATPase subcomplexes, the availability of structures of the intact enzyme in different rotational states enables comparison of 'open', 'tight', and 'loose' conformations of the AB pairs when the rotor is in different positions. The observed conformations of the A_3B_3 hexamer (Fig. 4a and Extended Data Fig. 6a–c) reveal a bend of the C-terminal domain of the A-subunit that closely resembles the conformational changes seen in the *Enterococcus hirae* V_1/A_1-ATPase[8] and mammalian mitochondrial F_1-ATPase[25] rather than the nearly rigid movement of subunits in the *T. thermophilus* V_1/A_1-ATPase[27]. The equivalent 'open', 'tight', and 'loose' conformations from different AB pairs can be overlaid with near perfect fidelity (Extended Data Fig. 6d–f). The protein samples used to prepare cryo-EM grids were not supplemented with nucleotide and consequently the nucleotide occupancy of the different catalytic sites is unknown. Because the nucleotide occupancy of the different AB pairs is not known, it is possible that the observed states correspond to the intrinsic asymmetry seen in AB pairs and αβ pairs of the *E. hirae* and *S. cerevisiae* V_1/A_1- and F_1-ATPase crystal structures lacking nucleotide, which closely resemble the conformations of the enzymes with bound nucleotide[8,28].

Rotary ATPases have been proposed to have an elastic coupling between their catalytic and membrane-bound regions to smooth the transmission of power between ATP hydrolysis or synthesis and proton translocation[3–5]. This need for elastic coupling is exacerbated by the 3:10 symmetry mismatch of the V_1 and V_O regions: the enzyme must deform to allow the rotor to be simultaneously in the correct position in the catalytic V_1 region and the membrane-bound V_O region. Earlier studies have suggested that the central rotor of the *E. coli* ATP synthase is the compliant element in that enzyme, while the peripheral stalk remains rigid[29]. The current structures indicate that the extended helical part of the central rotor D-subunit, equivalent to the F-type ATP synthase γ-subunit, remains rigid during rotation while the part of subunit D in contact with subunit d bends (Fig. 4b, dark blue, and Supplementary Videos 2 and 3). Further, the orientation of the d-subunit changes with respect to the D- and F-subunits, wobbling to accommodate distortion of the enzyme during rotation (Fig. 4b, cyan, and Supplementary Videos 2 and 3). The catalytic A- and B-subunits push against the E- and G-subunits of the peripheral stalks during rotation. The E- and G-subunits of the peripheral stalks undergo a bending motion along their elongated coiled-coil region, reminiscent of the action of a cantilever spring (Fig. 4c, Extended Data Fig. 6g–i and Supplementary Videos 2 and 3). The N-terminal domain of the a-subunit swings parallel to the membrane, moving to and away from the rotation axis of the rotor like the arm of a record player. The bend in the a-subunit occurs at the narrow interface between its N- and C-terminal domains (Fig. 4d and Supplementary Videos 2 and 3). In comparison, the 'head' domain of the C-subunit twists like a torsion spring at the 'neck' domain and thereby maintains its connections between peripheral stalks 2 and 3 (Fig. 4e and Supplementary Videos 2 and 3). The existence of these conformational changes is not obvious from inspection of crystal structures of the individual subunits. However, when visualized as a movie, each subunit appears to have evolved to perform these motions. Overall, the structures presented here show that in the V-ATPase both the rotor and stator parts of the engine undergo coordinated conformational changes. This combination of flexibility and rigidity is probably a key attribute for the function of these highly efficient macromolecular machines.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 November 2014; accepted 5 March 2015.

1. Sumner, J. P. *et al.* Regulation of plasma membrane V-ATPase activity by dissociation of peripheral subunits. *J. Biol. Chem.* **270,** 5649–5653 (1995).
2. Kane, P. M. Disassembly and reassembly of the yeast vacuolar H^+-ATPase *in vivo*. *J. Biol. Chem.* **270,** 17025–17032 (1995).
3. Pänke, O., Cherepanov, D. A., Gumbiowski, K., Engelbrecht, S. & Junge, W. Viscoelastic dynamics of actin filaments coupled to rotary F-ATPase: angular torque profile of the enzyme. *Biophys. J.* **81,** 1220–1233 (2001).
4. Stewart, A. G., Lee, L. K., Donohoe, M., Chaston, J. J. & Stock, D. The dynamic stator stalk of rotary ATPases. *Nature Commun.* **3,** 687 (2012).
5. Zhou, M. *et al.* Ion mobility–mass spectrometry of a rotary ATPase reveals ATP-induced reduction in conformational flexibility. *Nature Chem.* **6,** 208–215 (2014).
6. Walker, J. E. ATP synthesis by rotary catalysis (Nobel Lecture). *Angew. Chem. Int. Edn* **37,** 2309–2319 (1998).
7. Walker, J. E. Keilin Memorial Lecture. The ATP synthase: the understood, the uncertain and the unknown. *Biochem. Soc. Trans.* **41,** 1–16 (2013).
8. Arai, S. *et al.* Rotation mechanism of *Enterococcus hirae* V1-ATPase based on asymmetric crystal structures. *Nature* **493,** 703–707 (2013).
9. Benlekbir, S., Bueler, S. A. & Rubinstein, J. L. Structure of the vacuolar-type ATPase from *Saccharomyces cerevisiae* at 11-Å resolution. *Nature Struct. Mol. Biol.* **19,** 1356–1362 (2012).
10. Rawson, S. *et al.* Structure of the vacuolar H^+-ATPase rotary motor reveals new mechanistic insights. *Structure* **23,** 461–471 (2015).
11. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343,** 1485–1489 (2014).
12. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16,** 673–683 (2008).
13. Hirata, R., Graham, L. A., Takatsuki, A., Stevens, T. H. & Anraku, Y. VMA11 and VMA16 encode second and third proteolipid subunits of the *Saccharomyces cerevisiae* vacuolar membrane H^+-ATPase. *J. Biol. Chem.* **272,** 4795–4803 (1997).
14. Nishi, T., Kawasaki-Nishi, S. & Forgac, M. The first putative transmembrane segment of subunit c'' (Vma16p) of the yeast V-ATPase is not necessary for function. *J. Biol. Chem.* **278,** 5821–5827 (2003).
15. Matthies, D. *et al.* High-resolution structure and mechanism of an F/V-hybrid rotor ring in a Na^+-coupled ATP synthase. *Nature Commun.* **5,** 5286 (2014).
16. Nishi, T., Kawasaki-Nishi, S. & Forgac, M. Expression and localization of the mouse homologue of the yeast V-ATPase 21-kDa subunit c'' (Vma16p). *J. Biol. Chem.* **276,** 34122–34130 (2001).
17. Powell, B., Graham, L. A. & Stevens, T. H. Molecular characterization of the yeast vacuolar H^+-ATPase proton pore. *J. Biol. Chem.* **275,** 23654–23660 (2000).
18. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481,** 360–364 (2012).
19. Stock, D., Leslie, A. G. & Walker, J. E. Molecular architecture of the rotary motor in ATP synthase. *Science* **286,** 1700–1705 (1999).
20. Nicholls, D. G. & Ferguson, S. J. *Bioenergetics* 3rd edn, Ch. 3 (Academic, 2002).

21. Bueler, S. A. & Rubinstein, J. L. Vma9p need not be associated with the yeast V-ATPase for fully-coupled proton pumping activity *in vitro*. *Biochemistry* **54,** 853–858 (2015).
22. Toei, M., Toei, S. & Forgac, M. Definition of membrane topology and identification of residues important for transport in subunit a of the vacuolar ATPase. *J. Biol. Chem.* **286,** 35176–35186 (2011).
23. Junge, W. & Nelson, N. Structural biology. Nature's rotary electromotors. *Science* **308,** 642–644 (2005).
24. Lau, W. C. Y. & Rubinstein, J. L. Subnanometre-resolution structure of the intact *Thermus thermophilus* H^+-driven ATP synthase. *Nature* **481,** 214–218 (2012).
25. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. *Nature* **370,** 621–628 (1994).
26. Cingolani, G. & Duncan, T. M. Structure of the ATP synthase catalytic complex (F(1)) from *Escherichia coli* in an autoinhibited conformation. *Nature Struct. Mol. Biol.* **18,** 701–707 (2011).
27. Numoto, N., Hasegawa, Y., Takeda, K. & Miki, K. Inter-subunit interaction and quaternary rearrangement defined by the central stalk of prokaryotic V1-ATPase. *EMBO Rep.* **10,** 1228–1234 (2009).
28. Kabaleeswaran, V. *et al.* Asymmetric structure of the yeast F1 ATPase in the absence of bound nucleotides. *J. Biol. Chem.* **284,** 10546–10551 (2009).
29. Wächter, A. *et al.* Two rotary motors in F-ATP synthase are elastically coupled by a flexible rotor and a stiff stator stalk. *Proc. Natl Acad. Sci. USA* **108,** 3924–3929 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Rosenthal, R. Henderson, V. Kanelis, and L. Kay for comments on the manuscript. J.Z. was supported by a Doctoral Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada and a Mary Gertrude l'Anson Scholarship. J.L.R. is the Canada Research Chair in Electron Cryomicroscopy. This work was supported by operating grant MOP 81294 from the Canadian Institutes of Health Research.

Author Contributions S.B. and J.L.R. initiated the project. J.Z. and S.B. collected images and performed pre-processing steps. J.Z. performed the image analysis. J.Z. and J.L.R. interpreted the data, prepared figures, and wrote the manuscript. J.L.R. and J.Z. contributed new computer algorithms used in image analysis.

Author Information Cryo-EM maps have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-6284, EMD-6285, and EMD-6286. Atomic models have been deposited in the Protein Data Bank under accession numbers 3J9T, 3J9U, and 3J9V. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L.R. (john.rubinstein@utoronto.ca).

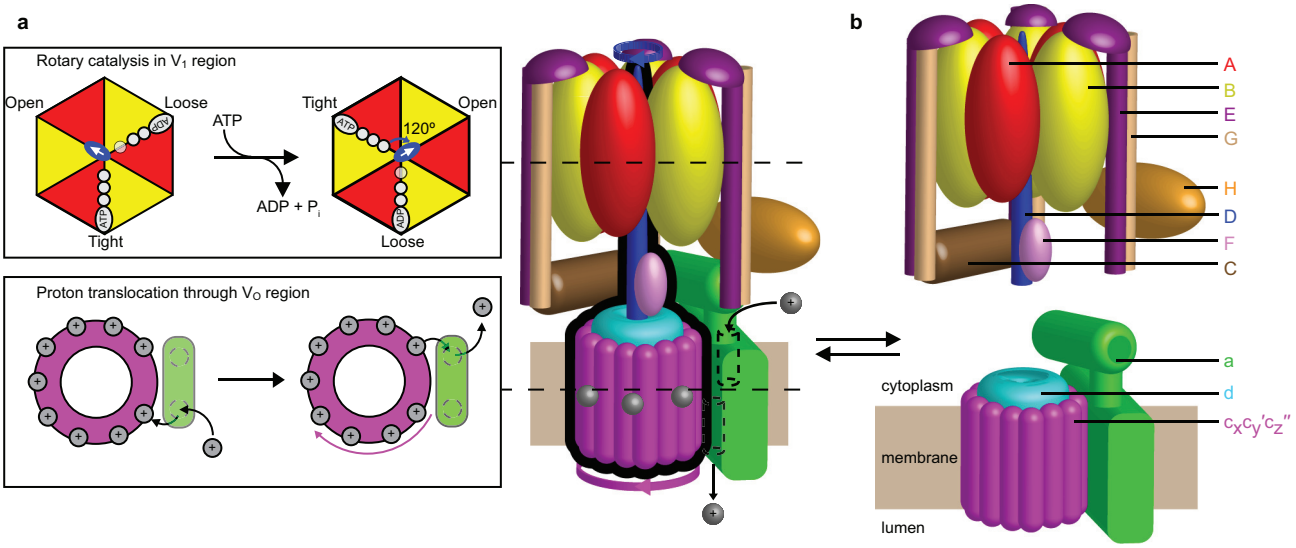
METHODS

No statistical methods were used to predetermine sample size.

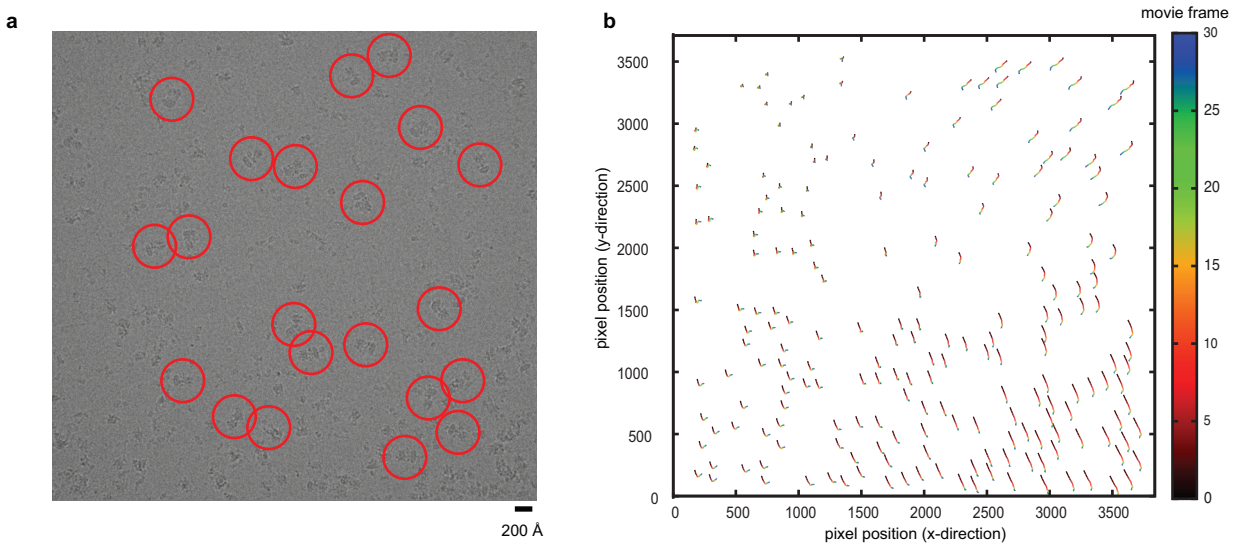
S. cerevisiae V-ATPase samples were purified as described previously from 10 l fermenter growths of *S. cerevisiae* strain JTY002 (ref. 9). The enzyme was purified by affinity to 3×Flag tags at the C termini of the A-subunits, providing samples (30 µl at ~10 mg ml^{-1}) in buffer containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 0.02% (w/v) DDM, and 150 µg ml^{-1} 3×Flag peptide, without any added nucleotide. Holey carbon-film-coated electron microscopy grids with regular arrays of 500–800 nm holes were prepared by nanofabrication[30] and cryo-EM specimens were prepared as described previously[9]. A total of 3,685 30-frame movies were acquired with defocuses between 1 and 7 µm with an FEI Tecnai F20 microscope operating at 200 kV and equipped with a Gatan K2 Summit direct detector device camera. The camera was used in counting movie mode with five electrons per pixel per second for 15 s and 0.5 s per frame. This exposure rate resulted in one electron per square ångström per frame on the specimen. Frames were aligned using the new program alignframes_lmbfgs[31] and averaged. Averaged frames were used for determination of contrast transfer function parameters with CTFFIND3 (ref. 32) and selection of coordinates for 200–250 particle-like features per image, some of which corresponded to true particle images, by template matching with TMaCS[33]. These coordinates were then used to extract candidate particle images from the unaligned movie frames with individual particle motion correction using the alignparts_lmbfgs algorithm[31]. A measured ~2% anisotropy in the magnification of the microscope was corrected in particle images by interpolation in Fourier space[34], to produce a calibrated pixel size of 1.45 Å. Contrast transfer function parameters were corrected to account for the effects of this anisotropy[34]. A total of 790,642 candidate particle images were subjected to two- and three-dimensional classification with Relion[35,36]. Particle images in two-dimensional classes with averages that resembled projections of the V-ATPase and contained high-resolution features were selected for further analysis. Three-dimensional classes were obtained from 217,533 particle images and three classes containing 106,445 particle images were used to build the three final three-dimensional maps. Local resolution was estimated with ResMap[37]. Calculations with Relion were performed using the SciNet cluster[38] and the SickKids High Performance Facility. Three-dimensional maps were segmented with UCSF Chimera[39,40]. Homology models were calculated either with Phyre2 (ref. 41) or with HHpred[42] and MODELLER[43] and atomic models from the PDB and homology models were docked into electron microscopy maps by MDFF[12]. Crystal structures were available for subunits C[44] and H[45], and the DF[46] and EG[47] subcomplexes. Homology models were built for subunits A, B[8,27], d[48], the N-terminal domain of subunit a[49], and the c-subunit[50]. The magnitudes of subunit rotations between states were measured in UCSF Chimera and movies were generated with UCSF Chimera.

Code availability. All new computer programs described above are available from https://sites.google.com/site/rubinsteingroup/home.

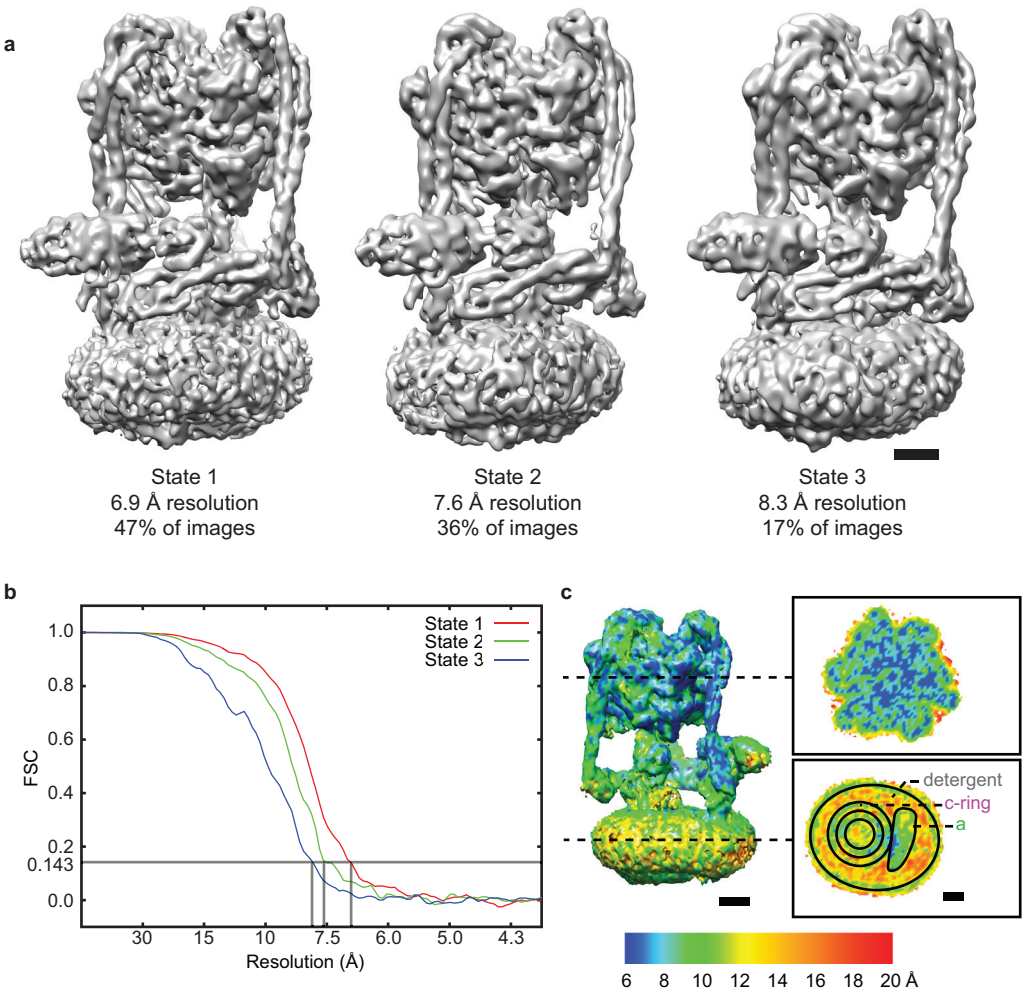
30. Marr, C. R., Benlekbir, S. & Rubinstein, J. L. Fabrication of carbon films with approximately 500 nm holes for cryo-EM with a direct detector device. *J. Struct. Biol.* **185,** 42–47 (2014).
31. Rubinstein, J. L. & Brubaker, M. A. Alignment of cryo-EM movies of individual particles by optimization of image translations. *ArXiv* **1409,** 1–11 (2014).
32. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
33. Zhao, J., Brubaker, M. A. & Rubinstein, J. L. TMaCS: a hybrid template matching and classification system for partially-automated particle selection. *J. Struct. Biol.* **181,** 234–242 (2013).
34. Zhao, J., Brubaker, M. A., Benlekbir, S. & Rubinstein, J. L. Description and comparison of algorithms for correcting anisotropic magnification in cryo-EM images. *ArXiv* **1501,** 1–10 (2015).
35. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
36. Scheres, S. H. W. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods* **4,** 27–29 (2007).
37. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
38. Loken, C. *et al.* SciNet: lessons learned from building a power-efficient top-20 system and data centre. *J. Phys. Conf. Ser.* **256,** 012026 (2010).
39. Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157,** 281–287 (2007).
40. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct Biol.* **170,** 427–438 (2010).
41. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4,** 363–371 (2009).
42. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** 244–248 (2005).
43. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
44. Drory, O., Frolow, F. & Nelson, N. Crystal structure of yeast V-ATPase subunit C reveals its stator function. *EMBO Rep.* **5,** 1148–1152 (2004).
45. Sagermann, M., Stevens, T. H. & Matthews, B. W. Crystal structure of the regulatory subunit H of the V-type ATPase of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **98,** 7134–7139 (2001).
46. Balakrishna, A. M., Basak, S., Manimekalai, M. S. S. & Gruber, G. Crystal structure of subunits D and F in complex give insight into energy transmission of the eukaryotic V-ATPase from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **290,** 3183–3196 (2015).
47. Oot, R. A., Huang, L. S., Berry, E. A. & Wilkens, S. Crystal structure of the yeast vacuolar ATPase heterotrimeric EGC(head) peripheral stalk complex. *Structure* **20,** 1881–1892 (2012).
48. Iwata, M. *et al.* Crystal structure of a central stalk subunit C and reversible association/dissociation of vacuole-type ATPase. *Proc. Natl Acad. Sci. USA* **101,** 59–64 (2004).
49. Srinivasan, S., Vyas, N. K., Baker, M. L. & Quiocho, F. A. Crystal structure of the cytoplasmic N-terminal domain of subunit I, a homolog of subunit a, of V-ATPase. *J. Mol. Biol.* **412,** 14–21 (2011).
50. Murata, T., Yamato, I., Kakinuma, Y., Leslie, A. G. & Walker, J. E. Structure of the rotor of the V-Type Na+-ATPase from *Enterococcus hirae*. *Science* **308,** 654–659 (2005).
51. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).



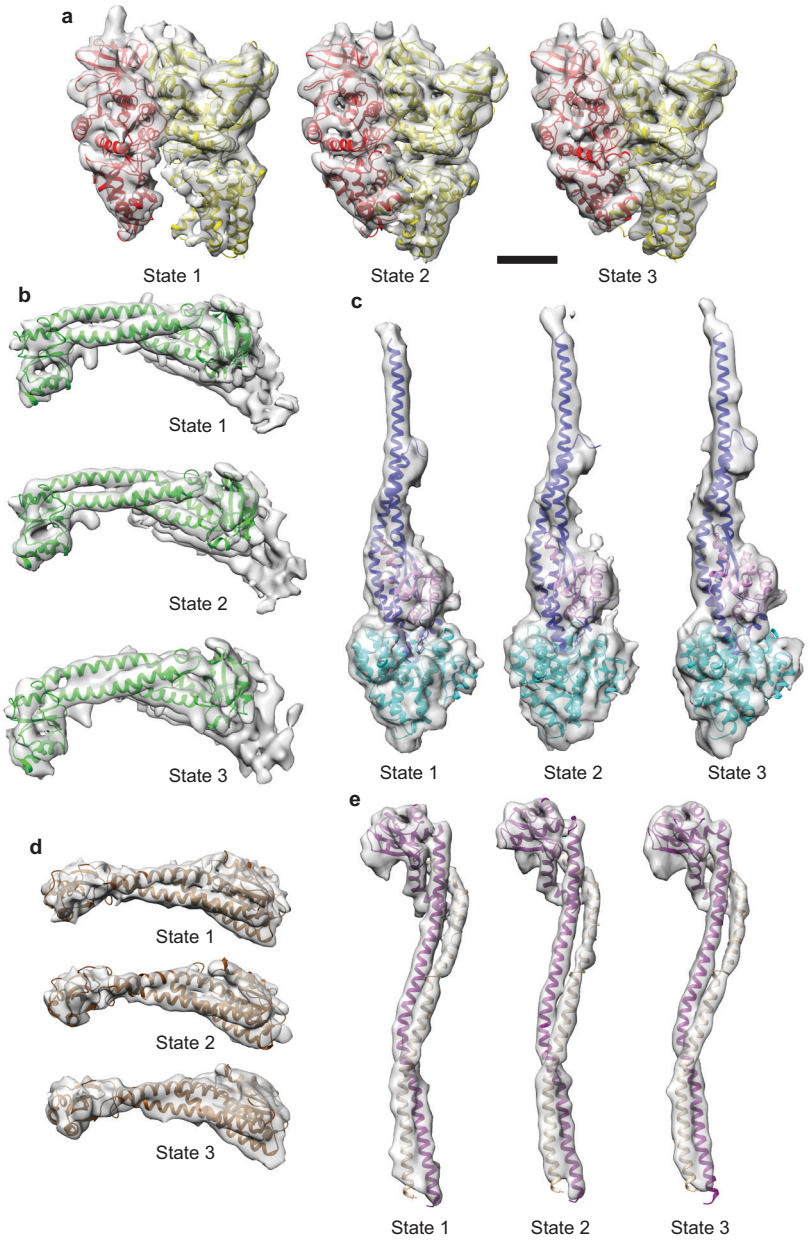
Extended Data Figure 1 | V-ATPase subunits and rotation. **a**, The V-ATPase from *S. cerevisiae* consists of subunits $A_3B_3CDE_3FG_3Hac_xc'_yc''_zde$, where *x*, *y*, and *z* denote unknown stoichiometries. Subunits with upper-case letter names correspond to components of the soluble V_1 region while lower-case names denote components of the membrane-bound V_O region. The e-subunit is not found in the detergent-solubilized *S. cerevisiae* V-ATPase. During rotary catalysis, ATP hydrolysis drives rotation of the rotor, consisting of subunits $DFc_xc'_yc''_zd$ (outlined in black), which rotates relative to the rest of the enzyme. Upper inset, the three different nucleotide-binding sites of the V_1 region can be found in three different conformations: 'tight' (where ATP is expected to be bound), 'loose' (where ADP is expected to be bound), and 'open' (where no nucleotide is bound). Lower inset, rotation of the $c_xc'_yc''_z$-ring against the a-subunit leads to proton translocation from the cytoplasmic side of the membrane to the luminal side of the membrane. Proton translocation occurs via two half-channels through the membrane. **b**, V-ATPase activity is regulated by reversible dissociation where the V_1 region separates from the V_O region. The H-subunit inhibits ATP hydrolysis in the dissociated V_1 region. Proton translocation in the dissociated V_O region is blocked by an unknown mechanism.



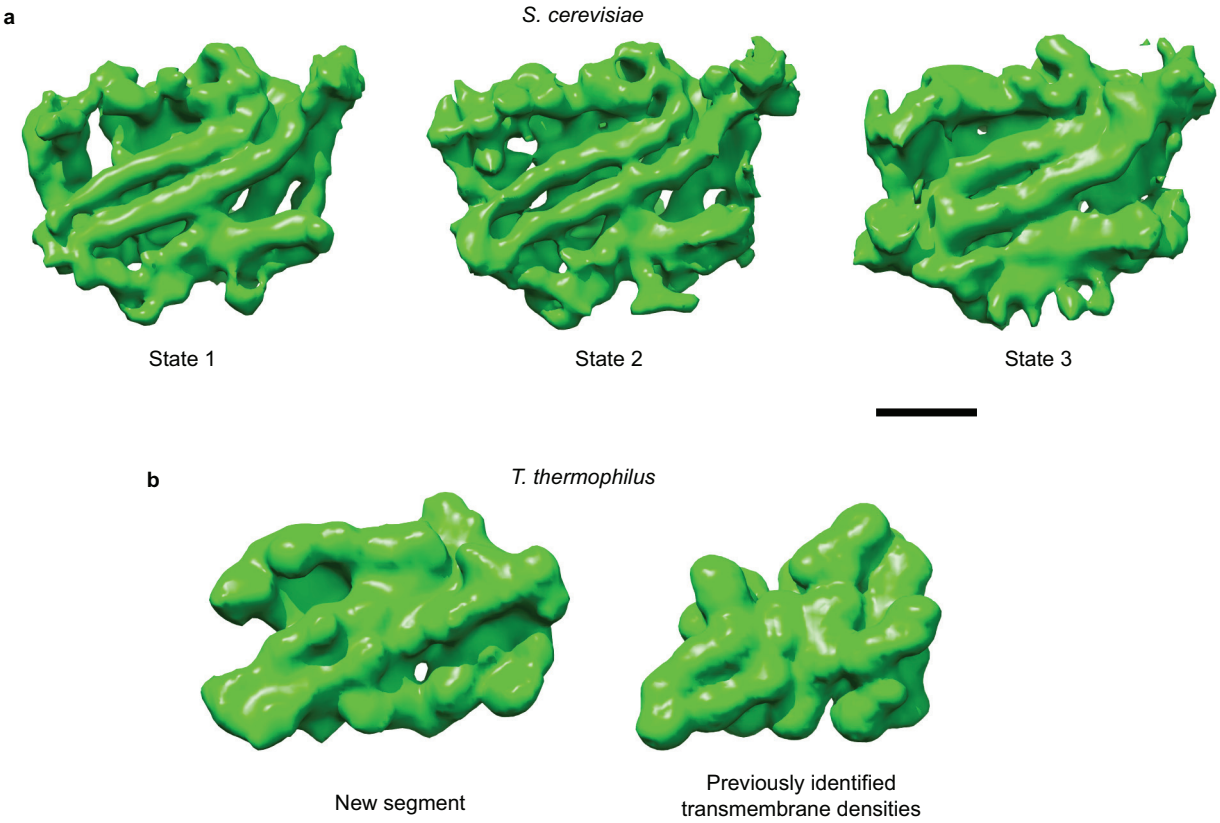
Extended Data Figure 2 | Data collection. **a**, A representative micrograph; examples of V-ATPase particle images are shown circled in red. These particle images were selected from the 200 candidate particle images identified automatically from template matching. **b**, Tracking of particle and other image feature trajectories with the alignparts_lmbfgs algorithm[31]. Trajectories are exaggerated by a factor of 5 to allow visualization.



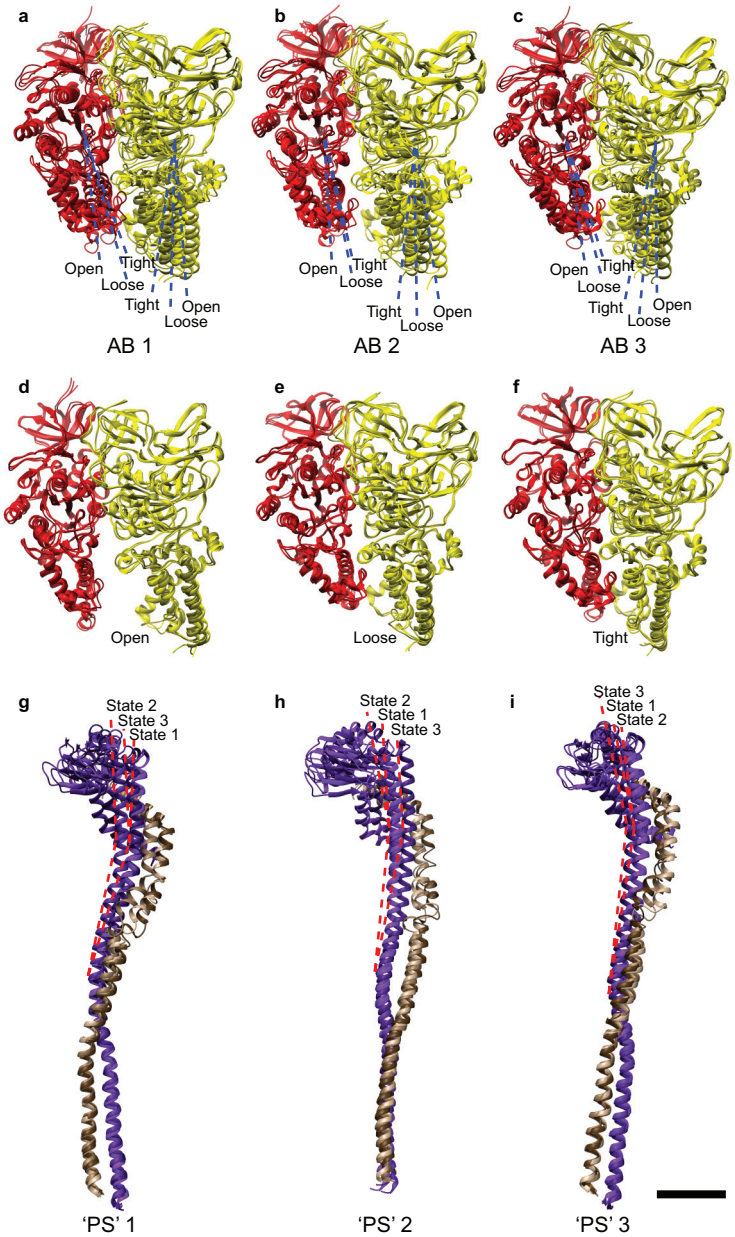
Extended Data Figure 3 | Three-dimensional maps from rotational states. **a**, Surface rendered views of the three three-dimensional maps are shown. Scale bars, 25 Å. **b**, Fourier shell correlation (FSC) curves after a 'gold standard' refinement of the three maps are shown. The resolutions measured from these curves at a Fourier shell correlation of 0.143 are the same as the resolutions measured after correcting for masking effects by high-resolution noise-substitution calculations[51]. **c**, Local resolution estimation shows that features in the V_1 region are better resolved than in the V_O region.



Extended Data Figure 4 | Map segmentation and molecular dynamics flexible fitting. Different subunits are shown fitted into their corresponding map densities in rotational states 1, 2, and 3, including AB pair 3 (**a**), the N-terminal domain of subunit a (**b**), the central rotor DFd subcomplex (**c**), subunit C (**d**), and peripheral stalk 1 (**e**). Scale bar, 25 Å.



Extended Data Figure 5 | C-terminal domain of the a-subunit. **a**, The membrane-bound C-terminal domain of the a-subunit appears similar in all three rotational states. **b**, The density from the C-terminal domain of the *Thermus thermophilus* subunit I, equivalent of the a-subunit, at 9.7 Å resolution[3] is consistent with the structure of the a-subunit from *S. cerevisiae* (left). However, the transmembrane α-helical densities identified previously in that map (right) are not entirely consistent with the current maps.



Extended Data Figure 6 | Flexibility in V-ATPase subunits. **a–c**, Each AB pair in the A_3B_3 hexamer goes through 'open', 'loose', and 'tight' conformations as the enzyme passes between the three rotational states. **d–f**, Overlay of all three open, all three loose, and all three tight structures shows that the conformations are nearly the same for each AB pair. **g–i**, Each of the three EG peripheral stalk structures undergoes similar bending motions between the three rotational states. Scale bar, 25 Å.

ERRATUM

doi:10.1038/nature14470

Erratum: Entanglement with negative Wigner function of almost 3,000 atoms heralded by one photon

Robert McConnell, Hao Zhang, Jiazhong Hu, Senka Ćuk & Vladan Vuletić

Nature **519,** 439–442 (2015); doi:10.1038/nature14293

In Fig. 2a of this Letter, the solid blue line in the linear plot was inadvertently removed during the production process; this figure has now been corrected in the online versions of the manuscript.

WORK–LIFE BALANCE Living is a journey, not a flow chart **go.nature.com/q34dzx**

ENTREPRENEURSHIP Scientists are experts at failing wisely **go.nature.com/reucok**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

MATEJ KASTELIC/SHUTTERSTOCK

ADJUNCT TEACHING

For love of the lecture

Contract teaching positions are becoming the norm for many aspiring professors. Know how to make the best out of them.

BY KENDALL POWELL

Last year, after finishing work each day at her administrative job with an art company, Susan Finley drove to Santa Monica College, California, to do what she loves most — teaching. On contract as an adjunct professor, she ran courses in psychology and counselling. She had 120 students last semester, but no office, which made holding office hours a challenge. Instead, she would meet students over dinner in the cafeteria or in a park on campus.

Contingent faculty members, also known as adjunct staff, can be full- or part-time, are off the tenure track and are generally paid on a per-course or hourly basis. As contract positions, the jobs bring greater challenges and less access to university resources than permanent positions. Yet despite the less-than-ideal working conditions, contingent faculty members such as Finley cherish the role for the tangible connection that it provides to a university and an academic department.

Before and after her classes, Finley advises and tutors students, a commitment for which she receives no compensation. "I'm one of the first professors these students have, and I love it," she says. "But I'm not paid for those hours, even though those are the hours I remember the most. Financially, it's not easy."

She is taking a gamble that the part-time work will help her to land a full-time teaching position — at least eventually. Thanks to other faculty members who have observed and reviewed her work, and students who have evaluated her performance positively, she ▶

has made valuable inroads on that goal. But there are no openings for a full-time faculty member in the department — and that is a burgeoning trend in the United States and other nations.

Although the number of part-time and contract teaching positions is growing quickly at universities across North America and elsewhere, full-time positions — both on and off the tenure track — have become more and more difficult to find. In 1975, tenured and tenure-track faculty members made up 45% of the faculty at US colleges and universities, and part-time instructors just 25%. By 2011, the numbers had reversed, with just 23% on the tenure track or already tenured, and part-time staff soaring to 42%. Today, more than two-thirds of instructional staff are non-tenure track, and most of them are on temporary contracts (see go.nature.com/gws6ow).

TEISHA J ROWLAND

Ethan Tsai delivers an organic-chemistry lecture at the Metropolitan State University of Denver.

WHET THE APPETITE

At 32 and with a newly minted doctorate, Finley thinks that teaching on a contract basis is a good way to test the waters of various career paths, but does not see it as a viable long-term prospect. Others in similar roles agree that this is the appropriate attitude to take in the current job market.

Non-tenure-track teaching positions, often called 'instructor' or 'lecturer', can offer freedom and flexibility, but often provide much-lower pay — salaries frequently start on a par with graduate-student stipends. For those who want to keep a hand in science, but not entirely at the bench, these roles offer a chance to hone teaching skills and work directly with undergraduate science students. They can also come with a hefty dose of advising and counselling responsibilities. Some instructors maintain limited research programmes; for others, research is out of reach.

Early-career scientists who are considering this route should understand that it is not generally realistic or sustainable as a long-term option or for those who want to engage in substantial research. Still, some instructors do manage to turn part-time posts into full-time positions, and a rare few jump to the tenure track at their institutions. Others use these posts as a bridge back to the academic environment following stints in other sectors, or as a way to supplement other work and fulfil their love of the classroom. Scientists who intend to land a full-time post in academic instruction will need to hone their time-management strategies and capitalize on their connections.

"I'm not paid for those hours, even though those are the hours I remember the most."

A passion for teaching is the most important attribute for success in the face of low pay, instructors say. Also desirable is a fondness for and an understanding of the university student population — knowing how to spark and maintain that group's attention. Beyond that, would-be instructors must be willing and able to invest large chunks of time for which they are not always compensated.

"I put a lot of thought into trying to keep students interested — what is this age group into?" says Pamela Buzas-Stephens, an adjunct geologist who juggles teaching positions at both the University of Colorado Boulder and Front Range Community College in Longmont. But determining how to keep students engaged is just the start, she says. Instructors need hours each week to prepare for a course, even if they have plenty of experience in front of a class. She estimates that an instructor who teaches a course for the first time will spend 3–6 hours preparing for every 1-hour session. In her experience, a 3-hour-per-week introductory course entails a minimum of 25 hours' worth of classroom activities, grading and advising.

There are ways to save time. For large introductory courses, first-time instructors could borrow notes, slides and syllabuses from colleagues who have taught it in the past. And they should prepare lectures far enough in advance to give themselves a buffer zone if they get busy with research or other obligations.

Buzas-Stephens relies heavily on allocating strict time slots for specific activities to limit the hours that she spends on course preparation. A PowerPoint presentation might get two hours, she says. Then it is time to move on.

Another huge time sink for many instructors is responding to student e-mails. Ken Diebel, an instructor in the fisheries and wildlife department at Oregon State University (OSU) in Corvallis, generates a 'frequently asked questions' document for each course to cut down on e-mails. Jennifer Stempien, a geology instructor and student adviser at the University of Colorado Boulder, has e-mail templates that she uses for common student dilemmas. She also preloads as much material as possible into the university's online course systems, then sets dates to roll out information as the semester progresses.

One of the trickiest aspects of Stempien's job — and a highly time-consuming one — has little to do with geology. Every year, she finds herself advising a handful of her 700 students on their daunting personal challenges, including physical and mental-health issues, and even suicide attempts and assaults. "I have all the appropriate campus offices on speed-dial," she says.

Technological advances such as the emergence of online instruction are putting extra strain on many instructors' time. Ecologist Luke Painter runs a mammal-systematics class online at OSU and had to adjust to recording video lectures. "I feel like I'm teaching in the dark, because I know the students are out there, but I can't talk to them directly," he says.

Diebel, who teaches an online course on riparian ecology, received a small grant last year for improvements such as videotaping lecture segments in the field. He attends an annual online-teaching symposium at OSU to learn what innovations other instructors use.

SIDELINE ACTIVITIES

With such heavy time constraints, it can be difficult to add research to adjunct jobs. But that has not stopped many instructors from trying. Stempien happened on an opportunity when she returned to the department in which she had done her postdoc to finish up a collaboration. While there, she ran into her department chair, who was looking for someone to teach a few introductory class sections. The chair hired her on a per-semester basis initially, and eventually offered her a three-year part-time

instructor contract. The move brought with it better benefits, compensation and access to departmental and university resources. It also gave her a place in the academic geology world that she loves.

She has struggled to carve out time to write up her research on student learning in the geosciences. Her work requires only database access, and she feels fortunate to be in a department that encourages her to continue her research. But it will have to wait until her summer break, when classes end.

Other instructors find ways to carry out at least some research during the academic year. Keen to continue her palaeontological studies on microfossils, Buzas-Stephens acquired a few necessary instruments and began to put in extra hours at her desk, examining specimens of tiny, shelled organisms called foraminifera. And Painter says that at OSU, instructors are allowed to apply for grant funding as investigators. He plans to follow this route to resume his research on the effects of bison and elk grazing on trees. US universities are increasingly offering this option to instructors to bring in more grants.

Others are happy to leave research out of the academic equation. Diebel, who worked for Oregon state agencies as a riparian specialist for 20 years, combines instructing with stream-restoration work as a private consultant. The consultancy allows him to teach part-time while maintaining his income, giving him more freedom and less bureaucracy. "The cool thing is, I don't have the 'publish or perish' or the tenure-promotion stuff to worry about. I focus on what I want to do, which is teaching," he says.

Many instructors select these part-time jobs for the flexibility, and may engage in consulting on the side, enjoy semi-retirement or spend time with family members. Some, however, view adjunct teaching as a stepping stone to a more-permanent post — for better or for worse. Teisha Rowland, a postdoc in cardiology at the University of Colorado Denver, is teaching general biology at the Metropolitan State University of Denver (Metro State) part-time to gain crucial teaching experience for her tenure-track job search.

"It's invigorating to be around that student energy the first time they learn about science," she says. But she is well aware that her adjunct position is best treated as temporary and as a way to build a CV. Buzas-Stephens agrees that science instructors who try to cobble together multiple low-paying adjunct contracts into a career are setting themselves up for heartache.

Andrew Robinson at Carlton University in Ottawa knows that heartache firsthand. He excels at the front of a lecture hall, but has struggled to convert his position teaching physics into a full-time, salaried post. "The chances of getting an academic job are already small — if you go into adjunct teaching, your chances probably get smaller."

Robinson and his wife both teach part-time to give them the flexibility to care for their son, who has special needs. But Robinson is bitter that, at age 52 and with 25 years of academic and industry experience, he works long hours teaching courses year-round to scrape together Can$35,000 (US$29,000) a year.

"There is no respect for contract instructors at all," says Robinson. "We are viewed as temporary replacements, which can be thrown away." His sentiments echo the frustrations of adjunct faculty members across North America, who have launched a movement to improve their working conditions and pay (see 'Groups give a voice to part-time faculty').

Still, scientists can take inspiration from the rare few adjunct instructors who do manage to make lemonade from lemons. Through hard work and perhaps a bit of luck, Ethan Tsai, a chemist at Metro State, achieved the near-impossible: he converted an instructor position to a tenure-track faculty spot in the same department.

Tsai began as a contract instructor for an organic-chemistry laboratory course and, over the next few years, was asked to take on more and more duties — including redesigning the lab-course curriculum and co-authoring a grant for the department to acquire a higher-resolution nuclear magnetic resonance (NMR) machine to analyse components of chemical samples. After the instrument arrived, he became its default manager.

Tsai ticked all the boxes for landing a tenure-track post: he made himself indispensable to the department by bringing in both grant money and research opportunities. He also showed his future colleagues that he was an enthusiastic collaborator who fit in the department's culture — a key aspect for tenure-track hires. Those connections made him the strongest candidate for his department's faculty search later on.

He describes his years on contract, when he was putting in 60-hour weeks, as doing heavy labour for graduate-student pay. Yet, he says, it was worth it for the reward of a tenure-track position — and his love of teaching made it possible. "If I didn't enjoy teaching, I would not have lasted beyond a year," he says. ■

Kendall Powell *is a freelance writer in Lafayette, Colorado.*

C. JERRY BOWEN

Adjunct Pamela Buzas-Stephens runs a field trip.

ADVOCACY FOR ADJUNCT ROLES

Groups give a voice to part-time faculty

Adjunct faculty members have connected across institutions, taking a stand against an education system built on jobs that are often precarious and demoralizing. Maria Maisto, president and executive director of the New Faculty Majority in Akron, Ohio, an advocacy group for US contingent and adjunct staff, says that the group formed in 2009 to provide guidance and support to the growing population of adjunct teachers. It aims to alert those inside and outside academia to the problems associated with such positions, by collecting statistics and advocating for practical solutions.

The Delphi Project, headquartered at the University of Southern California in Los Angeles, and the nationwide US Coalition on the Academic Workforce help adjunct faculty to organize and lobby for better working conditions. These groups support unionization efforts and work on legislative solutions at the state and federal levels. In 2013, Maisto testified before the US Congress for its 2014 report, *The Just-In-Time Professor,* which discusses the difficulties facing those in contingent roles (see go.nature.com/qtp9bs).

The Adjunct Project, a crowd-sourced database hosted by the *Chronicle of Higher Education* (adjunct.chronicle.com) in Washington DC, provides information on pay rates, benefits and working conditions across the United States. For example, the database shows that salaries for teaching biology in the Denver, Colorado, area range, at present, from US$950 to $8,000 per course. This kind of knowledge can help those in contingent roles to advocate for themselves. K.P.

WHEN LAST I SAW THE STARS

Lost vision.

BY JEFF HECHT

"I have news that no one else knows," I told my grandmother. "I promised I would tell you first."

"Yes, Little Helen. You have been good about that," she said. My parents called her Big Helen, because she had been so much taller when I was little. I grew just as tall as she was, but now she had shrunk and was too frail to stand for very long.

"Remember the new space telescope out past the Moon that is so big and so powerful it can see planets around other stars? I used it to see something new." Its 30-metre mirror sees far into the infrared, where planets like Earth glow with their own heat. I had told her before, but at 104 years old, she doesn't remember details well.

"You told me about that before, Little Helen. When last I saw the stars, you were just a little girl."

I was not so little then. I had been 12, visiting her and my grandfather at their summer cabin in rural Maine. We stayed up late, turned out all the lights, and went out to the open field behind the house. Leaning back in old aluminium lounge chairs, we saw the whole Universe spread across a deep black sky. My grandfather pointed out the Milky Way, handed me his big binoculars, and told me to look closely. When I focused the binoculars, the cloudy areas turned into more and more stars, and I was enchanted.

"They were beautiful," I told her. "I still remember that. It made me want to be an astronomer." Remembering how long ago that had been, I was grateful for the medical miracles that had let her live two decades longer than her grandparents had.

"Do you remember what I told you then, Little Helen?"

"About the Milky Way?"

"No, no," she said. "That we would never see the stars like that again. The first global geoengineering project was spreading haze across the sky to cool the ground. You didn't notice, but we could already see the scattered light in the night sky, as if city lights had come to our country cabin."

"We can see the stars better from space telescopes than we ever could from the ground," I told her. "There's no air in the way."

ILLUSTRATION BY JACEY

"I want to look up and see stars all across the sky, Little Helen," she said. "I haven't seen them in so many years; the haze and the lights are everywhere."

"We need the geoengineering aerosols, grandmother," I said. "They are the only way we can control the climate, to keep it from getting too warm, and melting the rest of Antarctica and Greenland."

"We could stop burning coal," she said. "I remember when governments talked about stopping greenhouse-gas emissions and not lighting up the roads and highways and the whole outdoors all night long. We called it conservation."

She and my grandfather had told me how they had worked to control pollution. They scared me when they told me what the future would bring. "Outdoor lighting does good things, grandmother," I said. "It lets us see when we go out at night. It prevents crime. It sustains our modern 24-hour lifestyle."

"There's too much light and too much pollution, Little Helen," she said. "But I don't want to argue. You said there was something you wanted to tell me. I want to hear your news."

I smiled. That was my grandmother, trying to avoid family arguments. "What I wanted to tell you was that the new telescope works so well we can see the disks of planets the size of Earth around other stars. We found one with oxygen in the air, and with other gases that are signs of life and civilization. We found the same sulfates that we spray into the stratosphere to keep our Earth cool. We think the planet may have a civilization." It was only 30 light years away, and I wondered why we had heard no signals from them. But I didn't want to complain to my grandmother that no one cared enough about other intelligent life to approve my plan to signal to them.

"That is wonderful to hear, Little Helen. I never thought that we would be able to see planets so far away. But what I want to see is stars sprinkled across a dark sky, and the whole glorious sweep of the Universe."

"I have pictures," I said. "I can show you."

"I don't want little pictures. You cannot know the Universe unless you see the stars spread across the whole sky," my grandmother said. "When I was young, I looked up at a sky full of stars and told your grandfather that someday our grandchildren's grandchildren would go visit the stars they saw in the sky. But if your grandchildren's grandchildren can't see the Universe of stars spread across the sky, they will never try to explore it."

I shivered, remembering how small my postcard of the Grand Canyon seemed after I had stood in the bottom of the canyon. I turned away so my grandmother would not see me cry. ■

Jeff Hecht *is Boston correspondent for* New Scientist *and a contributing editor to* Laser Focus World.

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

nature OUTLOOK

14 May 2015
Supplement to Nature Publishing Group journals

nature OUTLOOK

COLORECTAL CANCER

14 May 2015 / Vol 521 / Issue No 7551

Cover art: Mario Wagner

Colorectal cancer (CRC) kills almost 700,000 people every year, making it the world's fourth most deadly cancer (after lung, liver and stomach cancer). It is a disease of modernity: the highest rates of incidence are in developed countries (see page S2). As the world becomes richer, and more people shift to Western diets and lifestyles, the incidence of CRC is likely to increase.

But there is encouraging news. Screening for the disease has become routine in many parts of the world, and newer, less invasive technologies are being developed to replace the highly invasive colonoscopy (S4). And evidence is mounting that the disease could be largely prevented by a combination of drugs, nutritional supplements, diet and exercise (S6).

There is encouraging research on several fronts. The development of three-dimensional 'organoids' derived from adult stem cells could help match drugs to a patient's specific tumour type (S15). And DNA fragments that have escaped from tumours into the blood could be detected in 'liquid biopsies' that can characterize the tumour and help monitor the effectiveness of treatments (S9).

On the therapeutic front, however, progress is a matter of inches. New drugs have roughly doubled the average survival time for advanced CRC over the past decade, but the patient usually dies within three years (S12).

As with many other diseases, scientists are finding that the organisms that live inside us play a big role (S10). But whether an abnormal microbiome is a cause of CRC or an effect is still shrouded in mystery. This is just one of several fundamental questions about the disease that remain to be resolved (S16).

We are pleased to acknowledge that this Outlook was produced with the support of F. Hoffmann-La Roche Ag and Eli Lilly & Company. As always, *Nature* retains sole responsibility for all editorial content.

Herb Brody
Supplements Editor

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, Nature **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2015).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Colorectal Cancer* supplement can be found at http://www.nature.com/nature/outlook/colorectal-cancer
It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES
Feedback@nature.com

CONTENTS

RELATED RESEARCH

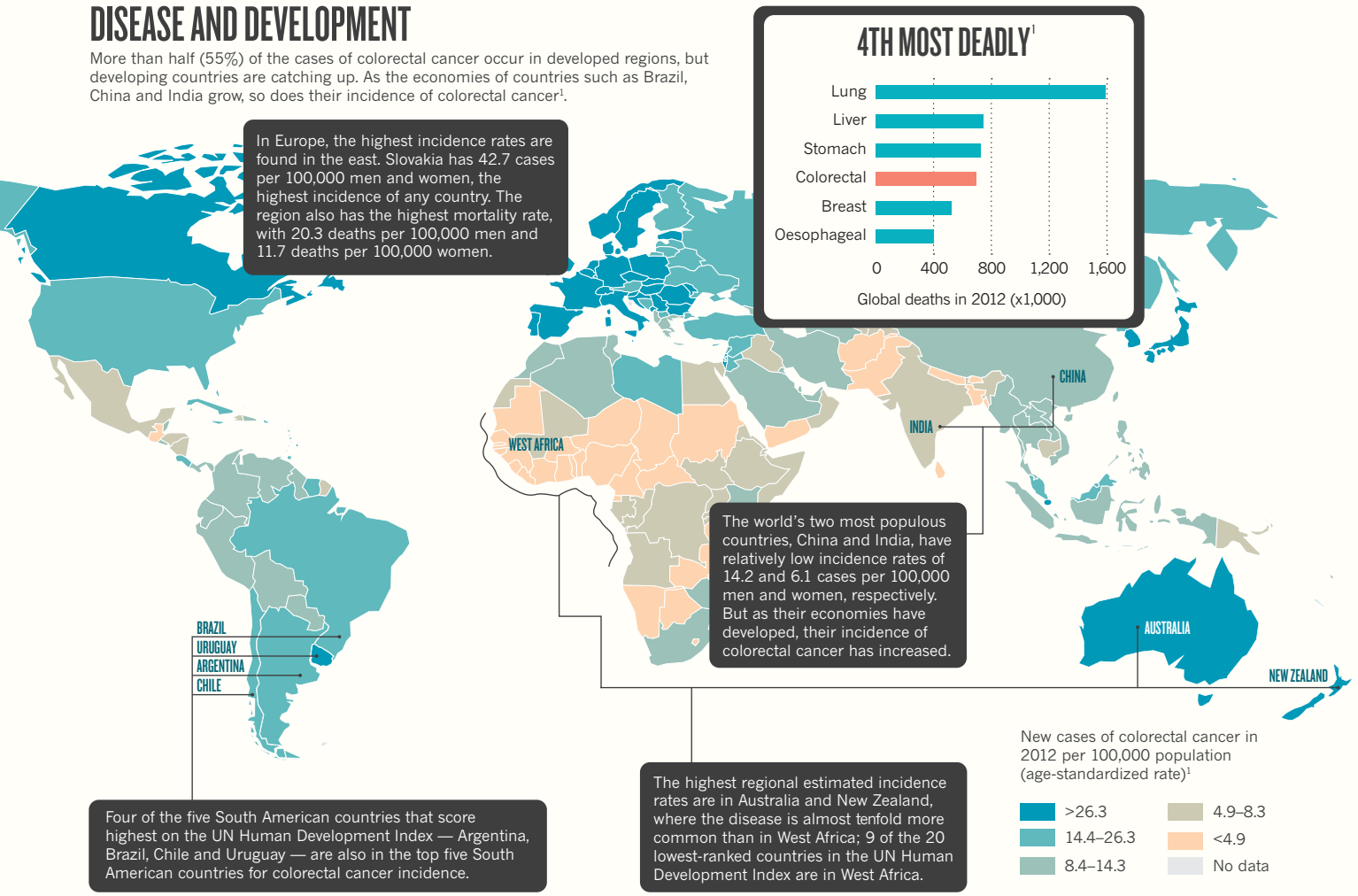
The latest papers on colorectal cancer published in *Nature*-affiliated journals.

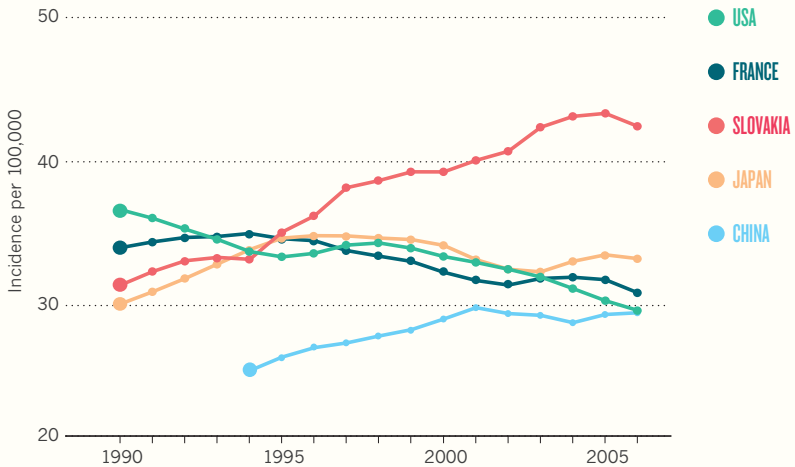
A DISEASE OF GROWTH

Colorectal cancer occurs throughout the world but is most common in developed countries. As heavily populated countries such as China undergo rapid economic development, the incidence of the disease looks set to increase. An animated version of this infographic is at go.nature.com/wgiqvp. By **David Holmes.**

DISEASE AND DEVELOPMENT

More than half (55%) of the cases of colorectal cancer occur in developed regions, but developing countries are catching up. As the economies of countries such as Brazil, China and India grow, so does their incidence of colorectal cancer[1].





HEALTH OF NATIONS

Global trends in colorectal cancer closely follow economic fortunes[2]. Incidence in Western Europe has been relatively flat for two decades, coinciding with a period of economic stability or decline. Meanwhile, Eastern European countries such as Slovakia have experienced rapid economic growth from a lower base, and have seen a corresponding rise in colorectal cancer.

Looking further east, Japan experienced a rapid rise in cases between 1990 and 1995, followed by almost ten years of steady or falling incidence. This period, termed 'the lost decade' by economists, saw falling wages and economic stagnation. By contrast, the breakneck development of Japan's neighbour China was accompanied by a dramatic rise in colorectal cancer incidence.

References: 1. Globocan, 2012 http://globocan.iarc.fr; 2. International Agency for Research on Cancer, http://ci5.iarc.fr; 3. Cancer Research UK; 4. Chan, D. S. M. *et al. PLoS One* **6**, e20456 (2011); 5. Larsson, S. C. & Wolk, A. *Int. J. Cancer* **119**, 2657–2664 (2006); 6. US National Cancer Institute Surveillance, Epidemiology, and End Results Program, http://seer.cancer.gov; 7. Galiatsatos, P. & Foulkes, W. D. *Am. J. Gastroenterol.* **101**, 385–398 (2006);

A PREVENTABLE EPIDEMIC

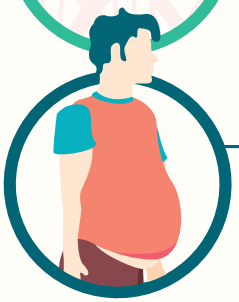
DIET

Many foods have been linked with increased (animal fat, sugars) or decreased (garlic, fibre, calcium) risk of colorectal cancer[3], but the most compelling evidence for increased risk relates to the consumption of red and processed meats[4,5].

GENETICS

Having a first-degree relative with colorectal cancer increases the risk of the disease by 80%. By the age of 40, almost all adults with familial adenomatous polyposis[7] will have colorectal cancer, and *BRCA1* mutations may also increase risk.

OBESITY



Closely linked to diet and physical inactivity, obesity also increases the risk of colorectal cancer. One 2013 study[10] found that obese people have a 33% higher risk of colorectal cancer than people of healthy weight.

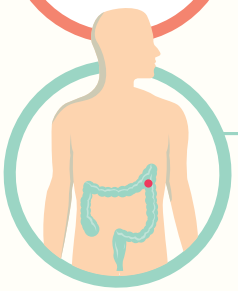
AGE

One of the most crucial factors for colorectal cancer is age. Up to 90% of all colorectal cancers occur in people aged 50 years and over[6].

EXERCISE

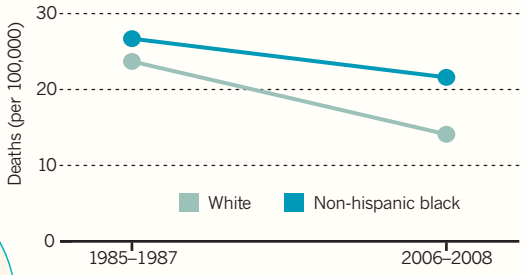
The World Cancer Research Fund and the American Institute for Cancer Research both identify exercise as protective against colorectal cancer. Colorectal cancer risk is 17–24% lower in the most physically active people compared with the least physically active[8,9].

SCREENING



Screening finds precancerous polyps that can be removed before they become cancerous. The problem is ensuring access. In the United States, for example, one-third of adults over 50 have not been screened appropriately.

SURVIVAL: NOT BLACK AND WHITE



The decline in colorectal cancer mortality in the United States has not been shared equally[11]. In the 1980s, African Americans died at a 13% higher rate than white Americans from the disease. Two decades later, they died at a 53% higher rate[12]. This disparity is caused by factors including differences in how likely they are to receive the latest treatments, and the prevalence of other health problems.

AMERICAN CANCER SOCIETY

HOW IT SPREADS

Colorectal cancer affects the colon and the rectum, parts of the digestive system known as the large intestine. It usually begins as a non-cancerous growth called a polyp on the inner lining of the intestine.

WHAT ARE THE STAGES OF COLORECTAL CANCER?

1. Cancerous cells form on the inner lining of the large intestine.
2. Cancer cells grow into the wall of the colon or lymph vessels.
3. Cancer cells penetrate the blood or the lymph vessels.
4. Cancer cells spread into nearby lymph nodes, and can also be carried in the blood vessels to other organs.

8. Robsahm, T. E. *et al. Eur. J. Cancer Prevent.* **22**, 492–505 (2013); 9. Boyle, T. *et al. J. Natl Cancer Inst.* **104**, 1548–1561 (2012); 10. Ma, Y. *et al. PLoS One* **8**, e53916 (2013); 11. Siegel, R., DeSantis, C. & Jemal, A. *CA Cancer J. Clin.* **64**, 104–117 (2014); 12. Robbins, A. S., Siegel, R. I. & Jemal, A. *J. Clin. Oncol.* **30**, 401–405 (2012).

AGENCJA FOTOGRAFICZNA CARO/ALAMY

Colonoscopy is often the best method for identifying colon cancer at an early stage — but it is invasive.

SCREENING

Early alert

Scientists are developing an array of choices for screening colorectal cancer, but patients often choose to go without.

BY CASSANDRA WILLYARD

Every day, a UPS truck delivers dozens of identical white boxes to the back door of a nondescript building in Madison, Wisconsin. Inside each box is a miniature pail filled with human excrement. This faecal matter is bound for Exact Sciences' 3,000-square-metre laboratory, a sparklingly clean facility that analyses each sample for the genetic markers of colorectal cancer.

The test, called Cologuard, was approved by the US Food and Drug Administration (FDA) in mid-2014. It comes at a time when public-health officials in the United States are desperately trying to get more people screened.

Only about two-thirds of those eligible for colorectal-cancer screening in the United States have been tested. In March 2014, a coalition of organizations dedicated to reducing colorectal-cancer deaths announced a target of 80% by 2018. Exact Sciences provides its at-home test as an alternative for people who are not willing to have a colonoscopy, an invasive procedure that requires sedation and a clean bowel.

Although colonoscopy is by far the most common screening option in the United States, researchers have yet to demonstrate its effectiveness in a randomized controlled trial. Other countries have embraced cheaper, less-invasive tests that have proved to be effective.

"Our goal is to be able to have the most people screened with an effective test," says Douglas Corley, a gastroenterologist at the US health-care provider Kaiser Permanente in Oakland, California. This might mean combining existing tests to create more effective screening programmes, or looking for new strategies that are not only effective but also appealing enough for people to get tested.

KING COLONOSCOPY

In the 1980s, colonoscopies were rare. Instead, the American Cancer Society recommended either sigmoidoscopy, a less-invasive procedure that uses a shorter scope to view only the bottom portion of the colon, or the faecal occult blood test (FOBT), which identifies blood in the stool — a potential sign of cancer.

The FOBT uses a chemical called guaiac to detect the haem component of the oxygen-carrying protein haemoglobin. Early FOBTs were poor at detecting malignancy, but more recent high-sensitivity versions detect between 50% and 79% of colorectal cancers that are found with colonoscopy. A variant of FOBT called a faecal immunochemical test (FIT), which uses an antibody to detect globin in the stool, can pick up between 55% and 100% of cancers detected with colonoscopy. But stool tests do not involve peering inside the colon and so are less effective at detecting precancerous polyps — a major drawback because detecting and removing these growths can prevent colorectal cancer from occurring in the first place.

Oncologist Alfred Neugut of Columbia University Medical Center in New York was one of the first to suggest, back in 1988, that colonoscopy might offer better cancer detection. Neugut argued that the long colonoscopy tube would allow physicians to screen a larger portion of the colon than was possible with sigmoidoscopy, so it would detect more cancers.

In 2000, US TV news personality Katie Couric, whose husband had died of colorectal cancer, encouraged her audience to have colonoscopies by undergoing the procedure on her show. In the same year, *The New England Journal of Medicine* published two studies[1,2] showing that colonoscopy can detect cancers missed by sigmoidoscopy. An editorial noted that "relying on flexible sigmoidoscopy is as clinically logical as performing mammography of one breast". The following year, federal health-insurance programme Medicare and many private insurers began paying for the procedure as a screening tool for colorectal cancer.

The US Preventive Services Task Force, an independent panel that issues evidence-based screening recommendations, lists three acceptable methods: FOBT alone, FOBT in combination with sigmoidoscopy, and colonoscopy. However, colonoscopy is the most common. According to 2012 data from the US Centers for Disease Control and Prevention, 65% of adults aged 50 to 75 reported being up to date with screening. The number screened by a colonoscopy within the past 10 years exceeded by more than 6-fold the number screened by FOBT in the past year, and by more than 60-fold the number screened with the FOBT–sigmoidoscopy combination.

If people could have only one screening test once in their lifetime, colonoscopy would be the clear winner, says Corley. But colorectal-cancer screening is not supposed to be a one-off event. The American Cancer Society recommends that people between the ages of 50 and 75 at average risk of the disease have either a colonoscopy every 10 years, a sigmoidoscopy every 5 years, or a stool test every year.

Colonoscopy brings a risk of bleeding and bowel perforation, and is less effective at catching cancer in the ascending part of the colon than in the descending portion.

NATURE.COM
For an animated overview of colorectal cancer, see *Nature Video*: go.nature.com/wgiqvp

The range of screening methods for colon cancer includes (left to right) stool blood test kits, faecal immunochemical tests and the Cologuard stool DNA test.

STEVE HORRELL/SPL; JAMES KING-HOLMES/SPL; EXACT SCIENCES

Three large, randomized trials have begun to study the value of colonoscopies, but they are unlikely to yield results for many years. By contrast, several trials have shown consistent results for FOBT and sigmoidoscopy. Sigmoidoscopy has been found to lower colorectal-cancer mortality by 22–31%, and FOBT reduces it by 15–33%.

Crucially, colonoscopies only work if people have them. In 2008, the US Preventive Services Task Force modelled the effectiveness of various testing strategies. With adherence at 80%, colonoscopy screening offered the greatest gain in life-years, followed closely by high-sensitivity FOBT and FIT. But when colonoscopy adherence was only 50%, it was no more effective than FIT, high-sensitivity FOBT, or either of those combined with sigmoidoscopy.

Other countries are keener to try the alternatives. In Australia, the National Bowel Cancer Screening Program sends FOBTs to people to use at home as they turn 50, 55, 60, 65, 70 and 74. Since the programme began in 2006, more than 4.6 million Australians have been offered the test, and 1.8 million have returned it for analysis. The goal is for all Australians between the ages of 50 and 74 to receive the test every two years by 2020.

The National Health Service (NHS) Bowel Cancer Screening Programme in England also favours FOBT. "It's recognized that colonoscopy is the gold standard," says Sally Benton, associate director of the NHS bowel-cancer screening hub in Guildford, UK. But the country lacks the resources to implement it.

In 2010, a team of UK researchers demonstrated the benefit of one-off sigmoidoscopy for colorectal-cancer screening[3], and the country is now running a pilot scheme for the procedure. Participants receive an invitation for sigmoidoscopy at the age of 55, and will receive FOBT kits every two years from the ages of 60 to 69.

DETECTING DNA

Cologuard provides another screening option. The stool test combines FIT with analysis to detect specific genetic markers — mutations in *KRAS*, a gene involved in cell division that is often mutated in colorectal cancer, and chemical modifications of two other genes associated with the disease. These markers provide a signature of the presence or absence of cancer, says David Ahlquist, a gastroenterologist at the Mayo Clinic in Rochester, Minnesota, who helped to develop Cologuard. "While many cancers and polyps don't bleed, they all shed cells."

Cologuard is not intended to replace colonoscopy — anyone who receives a positive result using any stool test still needs a colonoscopy to confirm it. But it does seem to have the edge over FIT. In a randomized trial[4] published in 2014, nearly 10,000 participants received either Cologuard or FIT in addition to a colonoscopy. Cologuard detected 92% of colorectal cancers compared with 74% for FIT, and found more precancerous lesions. But it also delivered more false positives: 13% of those who tested positive using Cologuard had no sign of cancer or precancerous lesions when they received a colonoscopy. And the study did not examine how the test would compare to FIT in the long term.

Douglas Rex, a gastroenterologist at Indiana University School of Medicine in Indianapolis, points out that about 70% of Cologuard's performance is due to its FIT component. What's more, stool DNA tests are recommended once every three years, whereas FIT is offered annually. If FIT were performed annually for three years, as it should be, "you probably would make up some of that difference", he says.

Then there is the cost. In 2014, Medicare agreed to pay US$500 for a Cologuard test — much higher than its $5 reimbursement for FOBT and $22 for FIT. Even if a person took the FIT test annually, as recommended, it would still cost much less than Cologuard over three years. "It's a little bit tricky to know whether the extra cost is worth it," Rex says.

BLOOD TESTS

To overcome people's queasiness about providing stool samples, several companies are developing tests that require nothing more than a drop of blood. "Many people don't want to touch the stool," says Ann Zauber, a biostatistician at the Memorial Sloan Kettering Cancer Center in New York.

German company Epigenomics markets one such blood test in Europe, called Epi proColon, which looks for a particular chemical tag on a gene called *SEPT9*. A study involving 8,000 people published in summer 2014 found that Epi proColon detected 68% of colorectal-cancer cases[5], but it had a high false-positive rate of nearly 20%. And, Rex points out, the test has a lower sensitivity for early-stage cancers than for late-stage cancers, which have a lower survival rate. For these reasons, he says, "it's not a good test for colon-cancer screening". A 2013 cost-effectiveness modelling study suggested that screening with Epi proColon every two years would be less effective and more costly than the alternatives.

But Epigenomics is targeting people who are not now being screened. "We want to lower the barrier for these patients to enter the screening programme," says Thomas Taapken, chief executive of the Berlin-based company. "Our assumption is that a blood test would do that."

In June 2014, the FDA declined to approve the test and asked the company to produce evidence that Epi proColon will increase compliance. Six months later, Epigenomics launched a study in the United States to answer that question. Researchers will invite people who have been offered screening but failed to comply to come into the clinic, where they will be selected at random to receive either a take-home FIT test or an Epi proColon blood test.

In a perfect world, adults would be screened for colorectal cancer when they are supposed to be, with a screening method that is proven to be effective. But in the real world, compliance is rarely perfect. Ultimately, says Zauber, "the best test is the one that gets done". ■

Cassandra Willyard *is a freelance science writer based in Madison, Wisconsin.*

1. Lieberman, D. A. *et al. N. Engl. J. Med.* **343,** 162–168 (2000).
2. Imperiale, T. F. *et al. N. Engl. J. Med.* **343,** 169–174 (2000).
3. Atkin, W. S. *et al. Lancet* **375,** 1624–1633 (2010).
4. Imperiale, T. F. *et al. N. Engl. J. Med.* **370,** 1287–1297 (2014).
5. Potter, N. T. *et al. Clin. Chem.* http://dx.doi.org/10.1373/clinchem.2013.221044 (2014).

MARIO WAGNER

PREVENTION

Tending the gut

Drugs, lifestyle changes and other measures might lower the risk of colorectal cancer — but the evidence is a long time coming.

BY LAUREN GRAVITZ

In a clinic at the University of Pittsburgh in Pennsylvania, gastroenterologist Robert Schoen has been injecting high-risk patients with a vaccine against colorectal cancer. With a series of three shots, he hopes to prime their immune systems so that they recognize, tag and destroy the disease in its earliest stages, before it can take hold and spread.

His trials[1] — the first to examine whether immunotherapy might be useful for preventing colorectal cancer as opposed to treating it — are part of a swell of research dedicated to cancer prevention that has taken hold in the past two decades.

"Prevention is the most important thing we can do from a public-health standpoint, but it's also the most difficult area for getting randomized proof," says Jason Zell, who studies the prevention and treatment of colorectal cancer at the University of California, Irvine. State-of-the art, double-blind, placebo-controlled trials are always difficult when it comes to disease prevention. And in the case of cancer, the aim is to thwart something that may not otherwise emerge for decades.

Colorectal cancer is largely preventable. Over the past 30 years, clinical trials and observational studies have suggested that everything from drugs to nutritional supplements, diet and exercise could help to stave off this lethal disease. The difficulties, however, lie in identifying what works, proving that it does, matching those people most at risk of cancer with the most appropriate interventions, and then persuading them to take the necessary action. And the whole process must be done in a way that is overwhelmingly safe.

Initially, many scientists believed that the best hope for preventing colorectal cancer lay in common elements of our diet. But even supplements that originate in everyday foods may not be as benign as they initially appear. Folate (from which folic acid is derived) may provide protection in the earliest stages of disease, but it actually seems to encourage the growth of tumour cells once they arise. And a large-scale trial aimed at preventing lung cancer showed that not only was β-carotene ineffective for prevention, it seemed to increase both risk of disease and risk of stroke.

Moreover, few scientists have the funds or the institutional structure to power the large and lengthy studies required to demonstrate effective prevention. "It takes a while for the progression of events to happen, and it's difficult to get a trial organized, completed and funded that can measure meaningful endpoints," says gastroenterologist Paul Limburg, director of preventive services at the Mayo Clinic in Rochester, Minnesota. Very few such studies have been designed for colorectal cancer, he says. To test prevention in a more realistic time span, most use the best proxy they can: decreased incidence of precancerous polyps called adenomas.

For now, the best way to prevent colorectal cancer is a colonoscopy (see 'Early alert', page S4). The cancer almost always starts as benign polyps that grow in the lining of the large intestine. The

➲ **NATURE.COM**
Visit *Nature Video* for an animation about colorectal cancer:
go.nature.com/wgiqvp

disease is so slow-growing that regular colonoscopies can prevent between 76% and 90% of malignant cases from developing into something more serious.

The problem is that not enough people are tested. In the United States, only about 65% of those for whom regular colonoscopies are recommended actually have them. For the rest, and even for those who are regularly screened, the addition of preventive therapy — an approach often referred to as chemoprevention — could save lives. "It makes sense to have something in addition to screening," says Limburg. "The ideal chemopreventive agent could delay screening or increase screening intervals."

Researchers are trying to figure out the most effective approach, but it is complicated — there may be many molecular pathways that trigger the cancer, and even more potential molecular targets. "The number of mechanisms we have is almost equal to the number of scientists investigating the problem," says Ernest Hawk, an oncologist and cancer-prevention researcher at the University of Texas MD Anderson Cancer Center in Houston.

INVESTIGATING INFLAMMATION

Despite all this complexity, colorectal-cancer researchers have been making progress. For colorectal cancer, "more than for any other cancer, there is evidence that there are agents we can use that would be effective", says Andrew Chan, a who studies colorectal-cancer prevention at Harvard Medical School in Boston, Massachusetts.

Early research suggested that the best way to stave off colorectal cancer was to inhibit enzymes that foster inflammation, such as prostaglandin and the cyclooxygenases COX-1 and COX-2, which encourage abnormal cell growth. The anti-inflammatory agent with the best long-term data so far is the humble aspirin.

Numerous studies over the past 20 years have shown that taking a daily dose of aspirin for a decade or more can reduce the risk of colorectal cancer by more than 50%, probably by inhibiting COX-1 and suppressing prostaglandin. But the effectiveness of aspirin may be limited by genetics. A study[2] published in March 2015 found that aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs) failed to protect a small percentage of people who had specific genetic variations on chromosome 15. In addition, taking aspirin in the long term can cause gastrointestinal bleeding, so oncologists recommend taking it only if a person's familial risk or history of polyps overwhelms the risk of bleeds.

Such side effects have led some groups to try to make aspirin safer for chronic use. At the University of Texas Health Science Center in Houston, Lenard Lichtenberger and his colleagues have found that lipids from membranes such as those lining the stomach and intestines guard the organs from injury. Aspirin eats away at that lining, so Lichtenberger has combined aspirin with lecithin — a natural phospholipid that helps to protect the mucous membrane without interfering with drug absorption.

The drug, developed for cardiovascular disease prevention by Houston-based PLx Pharma, was approved by the US Food and Drug Administration (FDA) in 2012 for the same uses as over-the-counter aspirin. The University of Texas team is conducting animal trials to evaluate whether it is as effective as aspirin at preventing colorectal cancer. The researchers hope to begin clinical trials in the next few years.

Beyond aspirin, the only therapeutics that have proven effective at lowering the risk of colorectal cancer are other NSAIDs, such as ibuprofen and COX-2 inhibitors. But whereas the evidence for aspirin is undisputed, these NSAIDS are on shakier ground. Their benefit is a bit lower — the reduction in disease risk is closer to 20–30% — and, like aspirin, most increase the risk of gastrointestinal bleeds and other side effects. Celecoxib, one of the best-studied of the group, decreases the risk of bleeding but increases the risk of cardiovascular complications such as heart attack and stroke.

One particularly promising NSAID is sulindac, an FDA-approved anti-inflammatory drug that not only inhibits COX enzymes but also helps to remove molecules called polyamines from the epithelial tissue that lines the colon and other organs. Elevated levels of polyamines have been associated with carcinogenesis, and inhibiting them has prevented the development of colon and other cancers in mice.

Zell and his colleagues are investigating the preventive power of sulindac combined with eflornithine, a drug that inhibits the creation of polyamines. In a three-year trial involving 375 people with a history of precancerous polyps, the combination of sulindac and eflornithine reduced the recurrence of adenomas by 70%, and progression to advanced adenomas by 92%, compared with placebo[3]. Now Zell is leading a trial of more than 1,300 people to find out whether the drugs can prevent the recurrence of colorectal cancer. By studying the two drugs in a four-arm trial — one placebo, one sulindac-only, one eflornithine only, and one both sulindac and eflornithine — Zell and his collaborators hope to find out whether the two drugs work better together than they do on their own. The trial is expected to end in 2019.

"We saw that 80% of colon cancer was really looking preventable."

LIVING PROOF

Lifestyle, too, has preventive potential. People who exercise regularly and who eat a diet low in red and processed meats have much lower rates of colorectal cancer. Yet although healthy habits are consistently associated with decreased disease, proving that these activities were the cause of lowered risk is difficult.

Most lifestyle studies are observational in nature — they use data from interviews or questionnaires to gather information on such factors as daily medication, diet and exercise, and then assess whether there is a link between behaviour and outcome. Graham Colditz, a cancer-prevention researcher and epidemiologist at Washington University in St. Louis, Missouri, conducted one such study[4] on the US population. "When we looked at national data comparing obesity, inactivity, red-meat and processed-meat intake, alcohol and cigarette smoking, we saw that 80% of colon cancer was really looking preventable," he says.

His study found that specific dietary factors seem to be implicated in a small fraction of colorectal cancers: about 6% can be attributed to eating red meat, and 5% to alcohol consumption. How they trigger carcinogenesis remains unclear. Some research suggests that the high iron content in red meat promotes carcinogenesis by increasing cell growth, and others indicate that one of alcohol's metabolites can interact directly with DNA. One recent study drives the point home. In a population of more than 77,000 people who refrained from smoking and drinking, those who ate fish but no other meat seemed to cut their risk of colorectal cancer nearly in half[5].

Obesity and physical inactivity also

JOHNER IMAGES/GETTY

Exercise can improve glucose metabolism and reduce the risk of developing colorectal cancer.

contribute significantly to colorectal-cancer risk, and here researchers have more confidence in the mechanism: they blame glucose metabolism. High levels of insulin, which the body uses to absorb glucose from the blood, seem to stimulate the growth of cancer cells, and the colon seems particularly sensitive to increased insulin. But exercise can halt or even reverse that cycle. The more physically active someone is, the more sensitive the body's cells become to the insulin, improving their ability to metabolize glucose. Moreover, contracting muscles use up more glucose. "We've looked at blood markers for insulin and seen that they relate to subsequent risk of colon cancer," Colditz says, "and exercise has a direct impact on glucose metabolism."

Among other research, data from the Harvard Nurses' Health Study — a long-term project that followed more than 200,000 nurses to assess how various factors influence health — found a strong connection between exercise and reduced risk of colorectal cancer, for both primary and recurring cancer.

The Canadian CHALLENGE trial, led by Kerry Courneya at the University of Alberta in Edmonton, is recruiting people without disease who have previously been treated for either stage II or stage III colorectal-cancer to find out whether they benefit from a physical-activity-based intervention. The researchers hope that this 6–8-year study will yield the most definitive answer yet about how exercise affects disease risk in colorectal-cancer survivors.

BUILDING A BETTER SHIELD

The link between insulin metabolism and cancer is prompting others to ask whether a medication that is already available could help to avert the disease. Michael Pollak, an oncologist at McGill University in Montreal, Canada, is testing metformin, a diabetes drug that suppresses excess glucose production. Metformin is one of the most commonly prescribed drugs in the West and has a long and proven safety record.

Its effects on colorectal-cancer prevention, however, are still unknown. What limited evidence there is comes from *in vitro* studies showing that, in the presence of metformin, epithelial cells from the colon proliferate more slowly — the opposite of what happens when cells start to turn cancerous. Metformin also has been shown to increase glucose uptake in the colon.

Pollak has confirmed the first finding in studies using human patients: taking metformin for a month significantly decreased the turnover of colon epithelial cells. But large-scale follow-up studies to validate the preventive effect are still a way off. "We certainly know that metformin has an excellent safety record — we have a million person-years of follow-up with it," Pollak says. "We would love to put together a prevention trial, but that's a big undertaking and people want more evidence."

"There is a lot of interest in diet and supplements for cancer prevention."

Two dietary supplements also raised hopes for prevention: calcium and vitamin D. But results have been mixed. Clinical epidemiologist John Baron at the University of North Carolina in Chapel Hill presented disappointing results at the American Association for Cancer Research conference in 2014. In a trial of more than 2,200 participants, he looked at the preventive potential of calcium and vitamin D, but found that neither of the supplements, nor a combination of the two, significantly reduced adenomas. "There are a lot of observational data suggesting that vitamin D would be dynamite," Baron says. "The results are a big disappointment." The same, he says, was true of calcium.

Such setbacks have left researchers feeling frustrated. "There is a lot of interest in diet and supplements for cancer prevention, particularly for cancers of the gastrointestinal tract," says Chan. "To date, however, there really are not any home runs."

Among all these unproven approaches to preventing colorectal cancer, the most intriguing may be the vaccine being developed at the University of Pittsburgh. Schoen believes that the immunotherapy approach that is gaining traction for cancer treatment — teaching a patient's immune system to recognize and attack antigens specific to cancer cells — can also stop colorectal cancer from developing in the first place. A vaccine given during the earliest stages of the disease, before adenomatous polyps have had the chance to progress into something more insidious, would prompt the immune system to develop an antibody against antigens expressed on precancerous cells.

Schoen's work focuses on a protein called MUC1, which is present on colorectal adenomas. In an early trial involving 40 people who had had advanced adenomas removed, the researchers administered three doses of the vaccine and measured the immune response. A year later, they gave the patients a booster. They found that nearly half of the subjects showed a significant immune response[1]. According to Schoen and his colleagues, this 'memory' means that if a vaccinated person develops an adenoma that expresses MUC1, the immune system could potentially recognize the protein and attack the cells expressing it, preventing cancer from developing.

The initial results were so promising that Schoen and his colleagues are now conducting a multicentre trial to determine whether the vaccine can prevent the recurrence of adenomas.

In the end, many researchers believe that the best option for preventing colorectal cancer might be a combination of several approaches. "Prevention may not be one-size-fits-all," says Limburg. "In fact, it probably isn't." But with the right mix — and unlike with other cancers, such as those of the pancreas or the blood — it may well be possible to prevent colorectal cancer. ■

Lauren Gravitz *is a freelance science writer based in Hershey, Pennsylvania.*

1. Kimura, T. *et al. Cancer Prevent. Res.* **6,** 18–26 (2013).
2. Nan, H. *et al. J. Am. Med. Assoc.* **313,** 1133–1142 (2015).
3. Meyskens, F. L. Jr *et al. Cancer Prevent. Res.* **1,** 32–38 (2008).
4. Joshu, C. E., Parmigiani, G., Colditz, G. A. & Platz, E. A. *Cancer Prevent. Res.* **5,** 138–145 (2012).
5. Orlich, M. J. *et al. JAMA Intern. Med.* http://dx.doi.org/10.1001/jamainternmed.2015.59 (2015).

JOHNS HOPKINS KIMMEL CANCER CENTER

Q&A Victor Velculescu

Out for blood

Oncologist Victor Velculescu, co-director of cancer biology at the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center in Baltimore, Maryland, describes how circulating tumour DNA can be used to improve the detection and treatment of colorectal cancer.

What is circulating tumour DNA (ctDNA) and how might it help in cancer therapy?
In ctDNA, fragments of DNA that escape from tumours into the bloodstream provide genetic clues that help us to detect and analyse the disease. These clues are especially valuable when tissue biopsies are not available, both at the start of treatment and when monitoring patients as therapy progresses.

Is ctDNA research still mainly in the lab?
We are moving to clinical applications. One application is looking at late-stage cancers. In many cases, tissue analysis is difficult or impossible, so a blood-based biopsy could be very helpful. We have shown in colorectal cancer that a liquid-biopsy approach allows us to identify patients with amplification of the HER2/neu receptor — a target of the drug trastuzumab (Herceptin). A small percentage of people with colon cancer have dramatic amplification of HER2/neu, and there are ongoing clinical trials that will figure out if it is a good target in colon cancer, and whether it will provide a therapeutic option for these patients.

Are you also looking at monitoring patients' responses to treatment?
Many patients are in clinical trials for a long time, and it may not be obvious whether they are responding to the investigational agent. A liquid biopsy makes it possible to get a measure of tumour burden that very nicely tracks disease over time. Clinical trials are extraordinarily expensive, and often we have to wait until the end to see if the therapy has worked. Monitoring subjects over the course of the trial by using liquid biopsies could provide insight into therapeutic responses much more quickly.

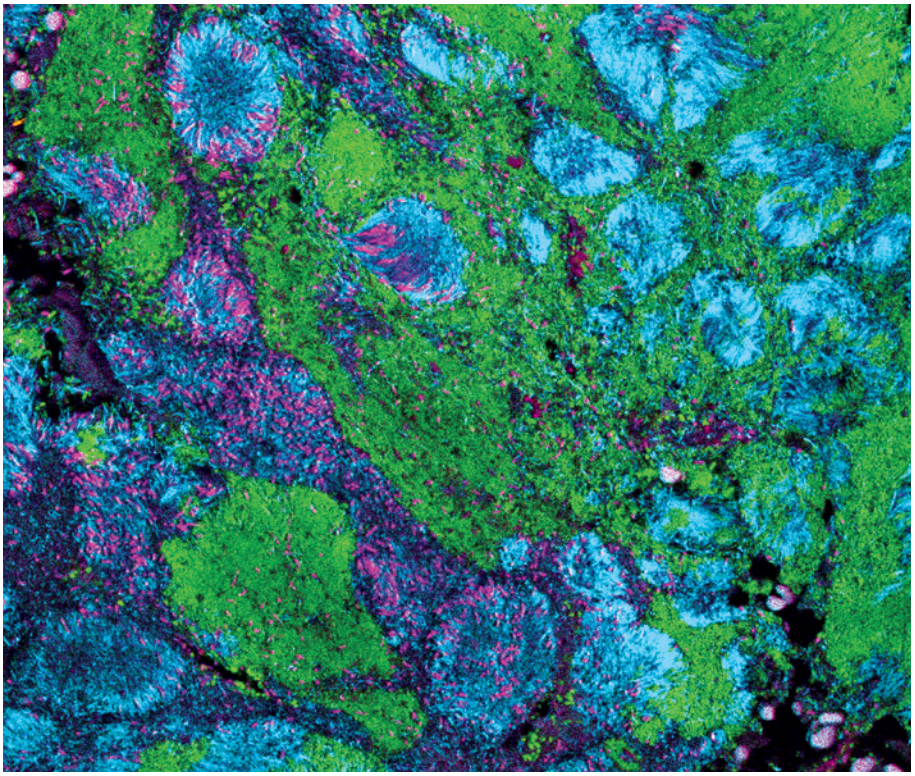
Can these liquid biopsies help us to understand resistance to drug therapy, by picking up clues about new mutations?
Yes, they allow us to study drug resistance, which arises from genetic alterations that occur in the tumour over time as it responds to the evolutionary pressure of the therapy. For example, we looked at a person with colon cancer who initially responded to inhibition of a protein called epidermal growth factor receptor (EGFR), but over time the tumour became resistant. We did a whole-genome analysis and found that the tumour had an amplification of the *MET* gene, which provides an alternative route for tumours to grow even in the presence of the EGFR blockade. In that case, we identified not only a mechanism of resistance but also a new potential avenue of therapy, because a number of *MET* inhibitors are now available or in clinical trials.

How can ctDNA aid the early detection of cancer?
One example is the early detection of disease recurrence. If you knew soon after surgery by using a liquid-biopsy test that the patient still had residual cancer, you could give some type of targeted therapy to help clear out the remaining cancer. Ultimately, liquid biopsies could make it practical to look at blood from apparently healthy individuals and try to determine whether they have a certain type of cancer. We should be cautious about this, though, because the levels of ctDNA in very-early-stage cancer might be below the limits of detection. But it will be an exciting avenue to investigate because a substantial fraction of patients with early, potentially curable disease are likely to be detectable in this way.

We are also seeing progress in the analysis of circulating tumour cells — cells shed from the tumour into the bloodstream that are still intact, as opposed to fragments of DNA. Are the roles of these two liquid-biopsy approaches becoming clear?
Yes, I think they are complementary. In some settings, ctDNA will be more sensitive for diagnostic purposes. On the other hand, circulating tumour cells are an extremely powerful source of tumour cells when you don't have access to the tumour itself. The challenge is to grow them and propagate them effectively outside the body. They could be a great avenue for testing new therapies.

Are drug companies starting to adopt ctDNA analyses?
There is a lot of interest from pharmaceutical companies, and liquid biopsies are beginning to be used as entry criteria for clinical trials. Companies are also interested in using liquid biopsies to monitor responses to therapy. If the analyses they do in their clinical trials are successful, companies may want to convert some of those tests into companion diagnostics. When they do, it is important for these diagnostic approaches to have the same rigour as the therapies. Like any approach, these liquid-biopsy genetic tests, and the regulatory framework guiding their use, won't be perfect. But we should not let the perfect become the enemy of the good. ■

INTERVIEW BY ERIC BENDER



A biofilm comprising three species of bacterium (green, cyan and magenta) on a colon tumour.

DEJEA, C. *ET AL. PROC. NATL ACAD. SCI. USA* **111**, 18321 (2014)

MICROBIOME

Microbial mystery

Gut bacteria have an important but elusive role in the formation of colorectal cancer.

BY SARAH DEWEERDT

Five years ago, four women and two men donated samples of their gut bacteria to science. Three women had healthy guts, but the other donors had colorectal cancer. Researchers implanted the human gut microorganisms into 'germ-free' mice, which were delivered by Caesarean section and reared in sterile cages, and so lacked gut microbes of their own.

This is a common technique for studying the health effects of the microbiome — the roughly 100-trillion bacteria that inhabit the gut. But the results were unexpected. When the scientists exposed the mice to a chemical mutagen that causes colorectal cancer, mice given microbes from colorectal-cancer patients developed fewer tumours than did mice that had received bacteria from healthy human donors[1]. "It was a very weird result," says study leader Patrick Schloss, a microbial ecologist at the University of Michigan in Ann Arbor.

The findings are emblematic of the puzzle researchers face when trying to understand the microbiome's contribution to colorectal cancer. "There's growing evidence for a role of bacteria in colon cancer, but it's not definitive," says Alberto Martin, an immunologist at the University of Toronto, Canada, who has studied microbial interactions with a biochemical pathway involved in an inherited form of colorectal cancer.

It makes intuitive sense that the gut microbiome has some role in colorectal cancer, whether for good or for ill. The idea is in line with evidence linking the microbiome to liver cancer and to other aspects of health and disease, such as obesity and the regulation of metabolism. Microbes are also known to be involved in the initiation and progression of inflammatory bowel disease, which is a risk factor for colorectal cancer.

What is more, most of the bacteria in the human gut make their home in the colon, or large intestine. "You almost never see cancer in the small bowel," says Jun Sun, a biochemist at Rush University in Chicago, Illinois. "Most bowel cancer originates in the colon." In other words, the cancer occurs where the bacteria reside.

There is more to the connection between the microbiome and colorectal cancer than theory and logic. Over the past decade or so, experimental evidence from animal studies, comparisons of gut bacteria from people with and without colorectal cancer, and analyses of the biochemical workings of individual bacterial strains all point to a link. But a full picture of how the microbial community in the gut acts and reacts during the development and progression of colorectal cancer remains elusive.

Many studies have documented differences between the microbiomes of people with colorectal cancer and people with healthy guts. The patterns vary widely from study to study, however, so scientists cannot say much about the microbiome abnormalities, or dysbiosis, in colorectal cancer. The most consistent finding is a high level of *Fusobacterium*, which is commonly found in the mouth but rarely in the healthy gut. Several studies have also shown that people with colorectal cancer have higher than normal levels of *Escherichia coli* bacteria[2].

But colorectal cancer can take decades to develop, so such differences may reveal little about the origin of the disease, says Christian Jobin, a microbiologist at the University of Florida in Gainesville. "What was cooking in their gut before the cancer diagnosis?" he asks.

Researchers have tracked changes in the microbiome throughout the course of colorectal-cancer development in mice, but there have been no such longitudinal studies in people. In 2014, however, Schloss and his colleagues reported differences between the microbiomes of three groups of people: those with healthy colons, those who had precancerous lesions called adenomas, and those with colorectal cancer. Adding microbiome data to conventional risk factors, such as age, race and body mass index, improved the ability to predict which of the three clinical categories a person belongs to — hinting at the potential to use microbiome analysis in colorectal-cancer screening.

But the cause-and-effect relationships still remain unclear. "What comes first — is it the cancer, the disease process, or the dysbiosis?" asks Shahid Umar, who studies gut bacteria at the University of Kansas in Kansas City. "That's the million-dollar question."

OF MICE AND BACTERIA

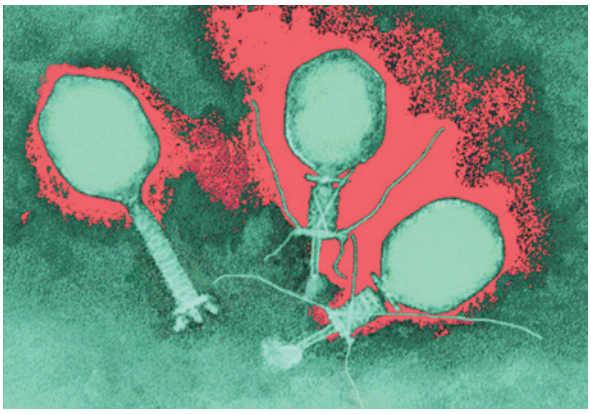
Direct evidence that gut-dwelling microbes help to cause colorectal cancer comes mainly from studies on mice. For example, Schloss and his colleagues transferred the gut microbiome from mice with colorectal cancer into germ-free mice, and then exposed these mice to a chemical mutagen. As long as this technique is used to transfer gut microbiota within a species, the results are straightforward: mice that received their microbiome from donors with cancer developed more and bigger tumours than mice that received their microbiome from healthy donor mice.

The researchers also showed that mice

exposed to the chemical mutagen do not develop tumours if they receive a dose of antibiotic that kills off some of the microbiome, suggesting that the microbiome plays a part in the onset of cancer. And they found that mice that receive a cancer-causing microbiome do not develop tumours if they are not exposed to the mutagen — so the mutagen is essential too. Other research groups have also found that cancer does not develop without the presence of certain bacteria, but that these bacteria do not, on their own, trigger tumours.

To unravel the biochemical mechanisms underlying these patterns, scientists have simplified matters by investigating individual bacterial species. This research has focused mainly on three types of bacterium: *Fusobacterium*; strains of *E. coli* that produce colibactin toxin; and toxin-producing strains of *Bacteroides fragilis* (a common cause of diarrhoeal illness). All three of these bacteria increase tumour formation in mouse models that are susceptible to colorectal cancer. For example, cancer-prone mice that are raised in a sterile environment and then colonized with *Fusobacterium* develop more tumours than mice not exposed to the bacterium.

LINDA STANNARD/UCT/SPL



Transmission electron micrograph of bacteriophages, which could be used to kill strains of bacteria implicated in colorectal cancer.

Not all species of bacteria have this carcinogenic effect, however. Jobin's team has shown that immune-compromised mice colonized with either *E. coli* or *Enterococcus faecalis* (another common gut species) develop colon inflammation, but only the mice that receive *E. coli* develop tumours. Inflammation is clearly important: transfer *E. coli* into a mouse that cannot produce an inflammatory response, and no tumours appear. It seems that an inflammatory environment in the gut changes the pattern of *E. coli* gene activity, Jobin says. To confirm this hypothesis, he has shown that deleting these inflammation-induced genes from *E. coli* removes the microbe's ability to cause cancer[3]. The carcinogenic effects of *B. fragilis* also depend on inflammation, according to Cynthia Sears and her colleagues at Johns Hopkins University in Baltimore, Maryland, who showed that the bacterium does not produce intestinal tumours when interleukin (IL)-17, an inflammatory molecule, is blocked[4]. Sears says that IL-17 is important in tumorigenesis in a variety of other body sites as well.

These animal studies paint a clear picture. "All this tells us that it's the microbiome," says Schloss. "And also importantly, it tells us that it's not just the microbiome."

HUMAN VARIATION

Demonstrating that similar processes occur in humans has been problematic, as illustrated by Schloss's study that found fewer tumours, not more, in mice that received microbiomes from human colorectal-cancer patients. But transferring microbiomes between species is an unnatural process and might be expected to throw up murky results.

Even so, there is some evidence that links *Fusobacterium*, *B. fragilis* and *E. coli* to colorectal cancer in humans. For example, *Fusobacterium* is more abundant in precancerous adenomas than in other sites in the colon. And the gene encoding *B. fragilis* toxin is present in the colons of 90% of people with colorectal cancer but in about half of the colons of healthy people. Sears points out that this is similar to observations for stomach cancer, which is associated with the bacterium *Helicobacter pylori*.

But it would be an oversimplification to think of the microbiome's role in colorectal cancer primarily in terms of a pathogenic infection by one or two troublesome microbes. "There are cases where individual bacteria seem to be associated with tumours," Schloss says. But far more often, he says, there is a "community effect" — a more generalized dysbiosis that may enable an individual cancer-promoting bacterium to gain a foothold, or to alter the microbiome's overall function in a way that leads to tumour formation.

Dysbiosis can have many different causes, but a common one is diet. High-fat Western diets have been linked to the development of colorectal cancer. Studies into the role of diet and fibre intake on colorectal cancer have yielded inconsistent results. Some researchers say this is because these studies have not taken into account people's different microbiota.

For example, individuals differ in their microbiome's capacity to produce butyrate, a fatty acid made by certain microbes when they break down dietary fibre. Butyrate is important for the health of the intestinal epithelium, which is one of the body's most rapidly renewing tissues — Umar points out that all the cells in the epithelium are replaced every 3–5 days.

Butyrate fuels this energy-intensive process and acts as a potent anti-inflammatory agent. It also feeds into important biochemical pathways, such as the Wnt and Notch pathways, which regulate epithelial-cell proliferation and have been implicated in colorectal cancer. "You need very controlled and regulated activity of these pathways," Umar says. Changes in microbial products — either the absence of butyrate or the presence of other metabolites — can affect these pathways and increase the proliferation of gut epithelial cells, a key step towards cancer.

Researchers should also consider the genetic background of individuals who have a particular community of microbes. Studies of mice with a genetic defect linked to a familial colorectal-cancer syndrome suggest that in individuals with this genetic background, butyrate can actually promote the development of colorectal cancer.

About 20% of colorectal cancers have similar genetic defects, which affect DNA-mismatch repair[5]. This means that it is crucial to carefully select the patients or subjects used in studies if we are to understand the effects of the microbiome on colorectal cancer, says Martin, who led the mismatch-repair work.

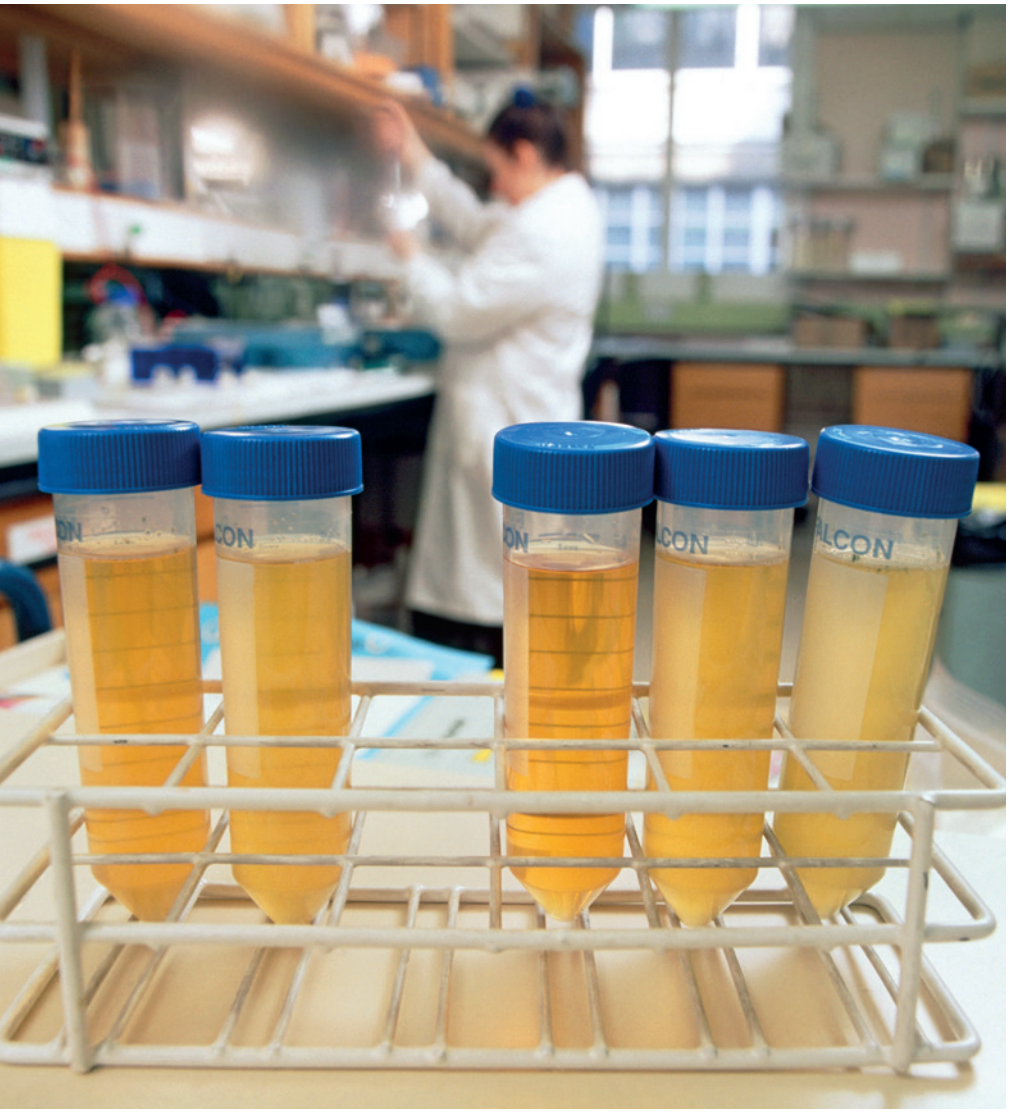
Unravelling all these effects is only the first step to understanding what causes colorectal cancer. "What we are after is not so much the nature of this dysbiosis, but whether we can do something about it," Jobin says. Some scientists have tentatively wondered about using antibiotics to knock out cancer-promoting gut microbes but worry about the potential effects on the microbiome as a whole. Jobin has a different idea: develop bacterium-killing viruses known as bacteriophages to target specific strains of bacteria that are implicated in cancer while leaving the rest of the microbiome to rumble along as usual.

Other scientists suggest that diet or probiotics (microorganisms thought to provide health benefits) could potentially manipulate the composition of the microbiome. However, this strategy needs a lot of work because researchers have had difficulty using probiotics to induce stable, long-term changes in the microbiome.

But microbiome interventions have shown promise for treating other gastrointestinal conditions. For instance, persistent infection with the bacterium *Clostridium difficile* responds to the unlikely intervention of faecal transplantation. Such examples offer hope that one day a microbial solution might also be found for colorectal cancer. ■

Sarah DeWeerdt *is a freelance science writer based in Seattle, Washington.*

1. Baxter, N. T., Zackular, J. P., Chen, G. Y. & Schloss, P. D. *Microbiome* **2,** 20 (2014).
2. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. IV & Schloss, P. D. *Cancer Prevent. Res.* **7,** 1112–1121 (2014).
3. Arthur, J. C. *et al. Nature Commun.* **5,** 4724 (2014).
4. Boleij, A. *et al. Clin. Infect. Dis.* **60,** 208–215 (2015).
5. Belcheva, A. *et al. Cell* **158,** 288–299 (2014).



JAMES KING-HOLMES/SPL

Faecal samples contain cells that can be tested for genetic mutations associated with colorectal cancer.

DRUG DEVELOPMENT

Mix and match

Doctors face a maze of drug options and genetic markers to find the right treatment for people with advanced colorectal cancer.

BY MEGAN SCUDELLARI

First, the good news: since 2004, the arsenal of approved drugs to fight colorectal cancer has more than doubled. Five targeted therapies newly approved by the US Food and Drug Administration (FDA) have joined four established chemotherapeutic agents that were already on pharmacy shelves. Together, these drugs have extended average overall survival for advanced colorectal cancer from 15 months to 30 months.

And the bad news? That figure for average survival has not climbed above 30 months. The new drugs have been of "marginal, incremental benefit", says Scott Kopetz, a colorectal-cancer specialist at the University of Texas MD Anderson Cancer Center in Houston. "They are not home runs." And a home run is badly needed, because colorectal cancer is the fourth leading cause of cancer-related death globally (third in the United States).

NATURE.COM
You can see an animation about colorectal cancer here: go.nature.com/wgiqvp

Surgery is used when colorectal cancer is diagnosed early and achieves an impressive 90% cure rate. But the other 10% of patients, and those for whom the disease was not detected early, develop advanced disease, marked by metastatic spread of tumour cells to the lymph nodes and other sites in the body. The 5-year survival rate plummets to just 11% for those patients whose disease has spread to distant organs.

Late-stage colorectal cancer is treated with drugs, usually a combination of general chemotherapy and a molecularly targeted therapy — a drug that interferes with a specific gene or molecule in a cancer cell. "In stage-4 disease, it takes an all-hands-on-deck approach," says Adam Snook, who develops treatments for colorectal cancer at Thomas Jefferson University in Philadelphia, Pennsylvania.

Targeted therapies give doctors more options for treating advanced colorectal cancer, but that very pool of choices provides the field's current challenge — figuring out which drugs to give to which patients. No two tumours are the same, and patients can have very different responses to the same drug, so clinicians find themselves involved in an elaborate guessing game to keep their patients alive. If scientists were able to link tumour characteristics, such as DNA mutations and gene-expression profiles, to positive drug responses, that would lead to diagnostic tests to select the best treatment for a given patient.

"We can now target virtually any kind of pathway and have the drugs available," says Thomas Seufferlein, director of internal medicine at Ulm University in Germany, who runs clinical trials for colorectal-cancer therapies. "Now we really need to match the treatment to the tumour."

Researchers are taking a three-pronged approach to try and overcome the 30-month plateau for colorectal-cancer survival. First, they are testing various combinations of approved therapies in clinical trials. Second, the recent identification of four molecular subtypes of colorectal cancer means that new biomarkers might help to identify which drugs are appropriate for individual patients. Finally, the effort to discover new drugs is continuing, with strategies that include targeting biochemical pathways in tumours that have so far resisted attack, as well as activating the immune system against the cancer.

DRUG ROULETTE

The first drug proven to work against colorectal cancer — 5-fluorouracil (5-FU) — arrived on the scene in 1962. For decades, it was the only effective colorectal-cancer drug. The late 1990s and early 2000s saw the development of three other chemotherapeutic agents, oxaliplatin, capecitabine and irinotecan, which all

inhibit DNA synthesis and stop cancer cells dividing (see 'Drug options'). When combined with 5-FU, these drugs made up the first chemotherapy combinations for colorectal cancer, with names that read like a strange game of Scrabble: FOLFOX, FOLFIRI and FOLFOXIRI.

Targeted therapies first appeared in 2004, starting with the monoclonal antibody bevacizumab, which inhibits the growth of new blood vessels. Bevacizumab was followed by a string of kinase inhibitors, which block the chemical messages that encourage cell division. "Each agent on its own has maybe not been that impressive, but combinations have greatly improved survival," says Snook.

"But," he quickly adds, "there's still a long way to go." That's because no single combination has successfully treated the wide diversity of tumours. DNA and RNA sequencing of cancer-cell genomes reveals that colorectal cancer is not a single disease, but rather many types of cancer that are instigated and propelled by different mutations in different people. There is no one-drug-fits-all solution. For now, combination therapy is the best option.

Part of the reason why combinations are necessary is the sheer amount of redundancy of cancerous pathways. Colorectal cancers, like many other cancers, can compensate for loss. If you target one pathway with an inhibitor, the tumour will become resistant by mutating and upregulating another pathway to perform a similar function as the blocked pathway. Unfortunately, using combinations of three or four drugs to block numerous pathways at once can become overwhelmingly toxic for patients. A phase III trial in Sweden, for example, found that general chemotherapy followed by a combination of two targeted therapies caused serious side effects in more than half of the patients, including bowel bleeding and holes in the walls of the intestine, with no difference in overall survival[1].

"Each agent on its own has maybe not been that impressive, but combinations have greatly improved survival."

So clinicians walk a fine line between trying to prevent tumour resistance and avoiding toxicity. The best way to navigate that line in the future, most agree, is to identify which patients will respond best to which combinations. That ability requires some sort of biological marker, typically a genetic mutation or a specific pattern of gene expression.

But colorectal cancer is too complex to have just one biomarker — researchers expect to find dozens. So far, though, only three biomarkers have been identified for advanced-colorectal-cancer drug treatment, and all are negative biomarkers, which means they indicate that a patient will not respond to a particular therapy. Patients with mutated *KRAS*, *NRAS* or *BRAF* genes do not respond to two of the available targeted therapies, panitumumab and cetuximab.

And that is the whole list. No other biomarkers exist to guide care for patients with advanced colorectal cancer.

DRUG OPTIONS

After decades of stagnation, the US Food and Drug Administration has approved several agents for metastatic colorectal cancer in recent years. The table shows all approved agents as of December 2013.

Class	Agent	Year	Characteristics
Cytoxic chemotherapy	5-fluorouracil	1962	Inhibits the action of an enzyme that synthesizes a nucleoside required for DNA replication; listed as one of the World Health Organization's (WHO's) 'essential medicines'.
	Capecitabine	2001	This 'prodrug' is administered in an inactive form; enzymes in the body convert it to 5-fluorouracil.
	Irinotecan	2000	Derived from the Asian 'happy tree', a deciduous tree native to southern China.
	Oxaliplatin	2004	A platinum-based drug that causes DNA to crosslink, preventing DNA synthesis.
Inhibits growth of new blood vessels	Bevacizumab	2004	A monoclonal antibody used to treat a variety of cancers; one of the WHO's essential medicines.
	Aflibercept	2012	First approved for the treatment of macular degeneration.
EGFR antibody	Cetuximab	2004	An antibody binds the EGFR receptor; it does not work if a downstream protein of EGFR, called KRAS, is mutated.
	Panitumumab	2006	Very similar in activity to cetuximab; a 2014 comparison found the two drugs to have similar overall survival benefit and toxicity.
Kinase inhibitor	Regorafenib	2013	The most recently approved drug for colorectal cancer, it is believed to inhibit at least eight different kinases.

Source: Sridharan, M. *et al. Oncology* **28,** 110–118 (2014).

FINDING SUBTYPES

Oncologists would love to find positive biomarkers, indicating that a patient will benefit from a particular therapy. If they could use molecular information to predict positive responses to treatment, they could turn the chaotic landscape of colorectal-cancer treatment into a smooth flowchart of expected outcomes. And they have a great deal of molecular information, thanks to the availability of next-generation sequencing technologies.

"It's like we stepped out of the age where you judge the weather by watching the birds," says Seufferlein. "This is what we did for a long time — judge the tumour by looking at X-rays or CT scans. Now, we just need to translate the molecular data into therapeutic strategies."

One of the first steps towards making this translation is to better understand the diversity of the tumours. That is why, three years ago, researchers around the globe began categorizing colorectal cancer into distinct subtypes based on gene-expression patterns. The first results came from The Cancer Genome Atlas Network in 2012, based on the analysis of 276 human colorectal-cancer tumours[2]. The study proposed three classes of colorectal-cancer tumour. Other teams subsequently proposed two, four, five, even six tumour subdivisions.

The jockeying to find the best classification was reaching a frenzy when, in a remarkable display of collaboration, more than 15 competing institutions joined forces to agree on the number of subtypes. Kopetz remembers being sceptical at the idea that so many differing groups, that had previously competing conclusions, might find consensus. But they did.

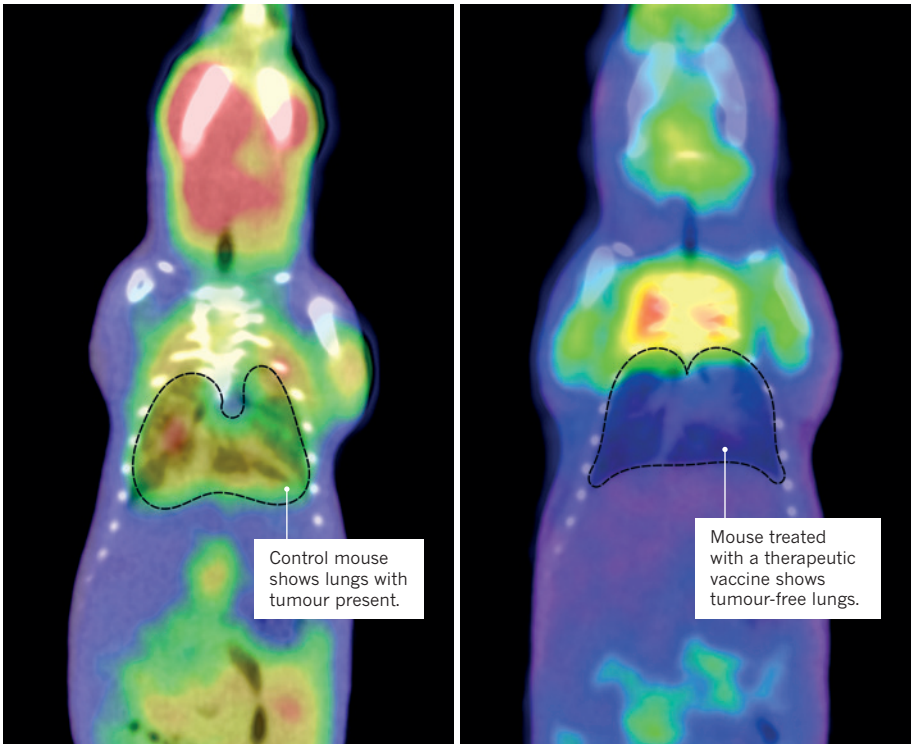
Organized by Sage Bionetworks, a non-profit company in Seattle, Washington, that specializes in collaborative data analysis, the Colorectal Cancer Subtyping Consortium used data from more than 5,000 tumour samples to identify four molecular subtypes[3]. "Everyone is setting aside their own classifications and saying this unified classification is what we should all use to move forward," says Kopetz. "It is tremendous."

So far, the consortium has identified some subtypes in which particular drug combinations clearly do not work, but many of their results are yet to be published. The goals now are to identify the best approved treatments for each subtype, and to use the subtypes to stratify patients in clinical trials.

DRUGGING THE 'UNDRUGGABLE'

There is still a huge need for new drugs to treat colorectal cancer, particularly therapies that are more effective and have fewer side effects than the current ones. But much of the pharmaceutical pipeline is made up of drugs that work in the same way as existing therapies. "There are a number of 'me-too' drugs, but we're not really tackling the root of colorectal cancer," says Federica Di Nicolantonio of the University of Turin in Italy, who specializes in preclinical drug research for colorectal cancer.

That root may be one, or both, of two types of mutation that pop up in every colorectal-cancer



ADAM SNOOK/SCOTT WALDMAN/THOMAS JEFFERSON UNIVERSITY

Immunotherapy has eliminated colorectal-cancer metastases from the lungs in mice (outlined area).

screen: *RAS* genes and the Wnt pathway. Neither has been successfully targeted with a drug, earning them the label 'undruggable'.

RAS mutations drive one-third of all human cancers, including 45% of colorectal cancers, by causing uncontrollable cell growth. For more than 30 years, scientists have tried to target the three human *RAS* genes — *KRAS*, *HRAS* and *NRAS* — but they keep hitting a brick wall. The RAS enzymes lack an active site that would normally be the target for a small-molecule drug, says Frank McCormick, who works on RAS at the University of California, San Francisco. And the rest of the protein's surface has no deep pockets that could make alternative targets. "It's just not a traditional enzyme," he says.

"We're trying to bring RAS back into focus as a drug target."

Because *RAS* mutations play such a large part in so many cancers, in 2013 the US National Cancer Institute began the RAS Initiative, a collaborative effort to better understand how *RAS* mutations drive cancer and to find new ways to silence the proteins. "We're trying to bring RAS back into focus as a drug target," says McCormick. "New technologies being developed are enabling us to target the protein in ways that were not possible when it was first attacked 15–20 years ago."

Academic teams are making progress with two small molecules that use chemical tricks to block RAS function. Both molecules are actively moving towards clinical trials in partnership with biopharmaceutical companies, says McCormick.

Less progress is being made in attempts to drug the Wnt pathway, a cascade of proteins that pass signals from the outside of a cell to the inside. This pathway is normally inactive in healthy adult cells but is activated by mutations in cancer cells. A 2012 analysis by the Cancer Genome Atlas Network found the Wnt pathway to be altered in 93% of all colorectal tumours[2]. Yet, like RAS proteins, the pathway has evaded all attempts to silence it.

A few Wnt inhibitors are now in early clinical trials, however. A team at the Icahn School of Medicine at Mount Sinai in New York, for example, is testing the addition of a soya-derived compound called genistein, which seems to inhibit a protein in the Wnt pathway, to a standard chemotherapy regimen in a phase I/II clinical trial. And drugmaker Novartis Pharmaceuticals is conducting a phase I trial of another potential small-molecule Wnt inhibitor. So far, there have been no late-stage clinical trials for Wnt inhibitors in colorectal cancer.

ACHIEVING IMMUNITY

As well as sending in molecules to kill tumour cells, scientists are also trying to activate the immune system to fight cancer cells. Immunotherapy has shown great promise in the treatment of other cancers, but has so far been unsuccessful against colorectal cancer.

One immunotherapy approach that showed remarkable results when treating melanoma and leukaemia, for example, proved disastrous against colorectal cancer. This 'adoptive cell therapy' technique involves removing immune T cells from a patient's blood, multiplying them in a dish (often after genetically engineering the cells to make them specifically attack tumour cells), and injecting them back into the patient. In a phase I trial that tested the procedure on three people with advanced colorectal cancer, all three experienced severe inflammation of the colon and had only temporary reductions in their tumours[4]. In another case, a person with advanced colorectal cancer treated with engineered T cells had a severe immune reaction and died[5].

Another type of immunotherapy deploys antibodies to shut down molecules that normally keep the immune system in check, unleashing the immune system to attack a tumour. But these 'checkpoint inhibitor' treatments have failed in colorectal cancer; in some studies, not one patient responded. "It's been a major disappointment," says Di Nicolantonio.

But not everyone has given up hope of using immunotherapy to treat colorectal cancer. Maybe drug developers have been using the wrong antigens — proteins on the surface of tumour cells that activate the immune system — says Snook, who is leading a clinical trial for a therapeutic vaccine.

Two antigens — CEA and MUC1 — have been widely studied in colorectal cancer and tested in therapeutic vaccines for more than 30 years, with mixed results. Snook's team recently turned to an antigen called GCC, which is found in 95% of colorectal-cancer tumours. The researchers are testing the safety and activity of a GCC therapeutic vaccine, and Snook hopes to have results by autumn 2015.

Immunotherapy may work on its own, or it could be combined with other treatments. Snook's team recently showed that radiation therapy followed by a therapeutic vaccine significantly amplified the immune response and shrank tumours in a mouse model of colorectal cancer[6]. "Neither therapy alone was particularly effective, but when we combined the two together, we saw a great benefit," says Snook.

Perhaps the future of colorectal-cancer therapy lies not only in combinations of drugs, but in combinations of different treatment types.

Until then, researchers continue to try to incorporate the newly defined colorectal-cancer subtypes and other biomarkers into clinical trials to improve the personalized treatment of advanced colorectal cancer. "We have all the molecular data available," says Seufferlein. "Now we need to translate it into therapeutic strategies." ■

Megan Scudellari *is a freelance science writer based in Boston, Massachusetts.*

1. Johnsson, A. *et al. Ann. Oncol.* **24,** 2335–2341 (2013).
2. The Cancer Genome Atlas Network *Nature* **487,** 330–337 (2012).
3. Dienstmann, R. *et al. Ann. Oncol.* **25,** ii115 (2014).
4. Parkhurst, M. R. *et al. Mol. Ther.* **19,** 620–626 (2011).
5. Morgan, R. A. *et al. Mol. Ther.* **18,** 843–851 (2010).
6. Witek, M. *et al. Int. J. Radiat. Oncol. Biol. Phys.* **88,** 1188–1195 (2014).

SANDER HEEZEN

Q&A Hans Clevers
Banking on organoids

In 2009, Hans Clevers and Toshiro Sato (then a postdoc in Clevers' lab) demonstrated a powerful new model to study development and disease: a three-dimensional 'organoid' derived from adult stem cells that replicates the structure of cells lining the intestine. More than 100 labs worldwide are now working with different types of organoid to study cancer and other diseases. Clevers, at the Hubrecht Institute in Utrecht, the Netherlands, discusses the potential of this approach.

Why might it be better to screen drugs in organoids rather than in cell lines?
We don't currently understand why certain tumours are sensitive or resistant to particular drugs. With targeted therapies, you can make a prediction, but for classical chemotherapy drugs, such as cisplatin or 5-fluorouracil, it is totally unpredictable which tumours will respond. Tumours can be sequenced in great detail, but drugs against them cannot be tested effectively other than in clinical trials. Organoids are a very good genetic representation of the tumour, so they let us bridge the gap between deep-sequencing efforts and patient outcomes.

How do you see organoids contributing to the study of colorectal cancer?
We are collaborating with groups at the Broad Institute in Cambridge, Massachusetts, and the Sanger Institute in Hinxton, UK, to build a biobank of organoids from 20 or so people with colon cancer. We have organoids of the cancer and of normal cells from individual patients, as well as sequences of their protein-coding genes. We have established the non-profit Hubrecht Organoid Technology (HUB) to expand our organoid biobanks. The HUB shares these biobanks with academic groups around the world, and now works with about 15 companies on drug-development programmes. We can culture tumours from almost every person with colon cancer, sequence them and test them against drugs. Additionally, we can use research techniques that have been developed for cell lines, such as genetic tools, fluorescence-activated cell sorting and microarrays.

Is this research moving towards clinical trials?
Yes, my group and the HUB are collaborating with Emile Voest at the Netherlands Cancer Institute in Amsterdam on an observational trial. We already have some organoid models from people with colon cancer who receive chemotherapy. The organoids are screened against a panel of common colon-cancer drugs. The patients will be treated the same way the oncologists would normally treat them, but we'll see if we could have predicted the response from our organoids. We're also starting another trial in which we will enrol advanced-colon-cancer patients, for whom there is no standard treatment. We will make organoids, test drug sensitivity and resistance, and then advise the oncologists as to what drug to use for that particular patient. We will be looking at multiple drugs, so we need large numbers of patients — that's the only way we will be able to produce enough data to help us match drugs to tumour types.

To benefit individual patients, won't you need to test the drugs very quickly?
Yes — and that's really where we want to take this technology. When you have pneumonia, your bacterial cultures are tested and you get answers in three days. With this technology, we can tell the oncologist the best odds for a combination of therapeutics, maybe not in three days, but in several weeks. We have an organoid-based test in cystic fibrosis that gives us a result in about two weeks.

How does the organoid approach differ from patient-derived xenografts, in which patients' tumours are transplanted into immune-suppressed mice for testing drugs?
It's the same principle — you get a functional readout of the patient's tumour. But organoids can be tested against an unlimited amount of compounds and combinations. Furthermore, in contrast to xenografts, organoids can be established from almost all patients.

What are some of the next steps in your cancer research?
Organoids model the key component of the tumour but they lack some important elements. We want to combine organoids with other elements to make more-complete tools. For instance, we would like to introduce the immune system so that we can study the effects of the fantastic new immunotherapy drugs. We think that we can build it up in a reductionist way — take lymphocytes isolated from a tumour, bring these together with cancer organoids derived from the same tumour and watch what happens. And maybe we can also put microorganisms in these organoids. For example, we could add *Helicobacter*, a major cause of stomach cancer, to stomach organoids.

Can organoids also help to test drug combinations?
Yes, tumours are genetically heterogeneous, and there can be vast differences in drug sensitivity between clones for the same tumour. We can possibly advance sequence-based therapy by testing millions of drug combinations in organoids. ■

INTERVIEW BY ERIC BENDER

MARIO WAGNER

COLORECTAL CANCER

5 BIG QUESTIONS

Research is attacking colorectal cancer on many fronts, with varying degrees of success. But solving these five central puzzles is likely to be crucial.

BY SHRADDHA CHAKRADHAR

	QUESTION	WHY IT IS HARD TO ANSWER	CURRENT UNDERSTANDING	QUOTE
1	***What drives the mutations that cause polyp formations in non-hereditary cancers?***	As with other multifactorial conditions, there are too many moving parts to make it easy to pick out specific culprits.	Some 5–10% of colorectal cancers are caused by heritable mutations. Scientists think that about 50% of the rest are caused by sporadic mutations.	"We know that mutations cause polyps, but we don't know what proportion of mutations are caused by replicative mutation and by environmental factors." **Bert Vogelstein**, Johns Hopkins.
2	***What determines whether a polyp will become cancerous?***	Researchers must tease out the sequence of mutation events that drives this transformation.	Larger polyps are more likely to become cancerous. Adenomas, one of the main types of polyp, lead to 95% of all colorectal cancers, but fewer than 10% of adenomas become cancerous.	"A lot of the information we have is from [people with] genetic abnormalities, and that only helps us infer what happens in normal people." **Neil Hyman**, University of Chicago Medical School.
3	***What environmental factors contribute to the progression of disease?***	Isolating crucial factors and combinations is difficult. Clinical trials take many years.	Sedentary lifestyles and diets that are high in fat and red meat are associated with an increased risk.	"All the empirical studies that have tried to determine what about the diet causes cancer haven't really been that successful." **Heidi Nelson**, Mayo Clinic.
4	***What makes some cancers more responsive to certain therapies than others?***	Colorectal cancer shows a high degree of heterogeneity in mutations, both between and within individual tumours.	Precise molecular profiles of tumours will help to predict the response to targeted therapies or combinations.	"We're throwing the dice — we really don't know what's going to work and what isn't." **Vogelstein**.
5	***How can immunotherapy be harnessed to treat colorectal cancers?***	We lack good ways of studying colorectal cancer in humans, and animal models (mostly mice) are inadequate. Some immunotherapies are in trials, but any clinical impact is years away.	Some tumours have infiltrating lymphocytes, which are white blood cells that can aid in killing the tumour.	"Every patient who shows up at my door wants to know if we can activate their immune system to kill cancer." **Charles Fuchs**, Dana-Farber Cancer Institute.

Shraddha Chakradhar *is a news editor at* Nature Medicine.